NOTE

This Microfiche meets the quality requirements of the American National Standard for Microfiche of Documents (ANSI PH5.9-1975). However, because of imperfections in the document from which it was made, it contains images that are illegible or difficult to read. The source document for this microfiche is the best copy that could be obtained prior to filming.

CIS

Congressional Information Service, Inc.
Washington, D.C.

Congress of the United States
Office of Technology Assessment

MAPPING OUR GENES
Federal Genome Projects: How Vast? How Fast?

Contractor Reports, Volume 1

February 1988

Bibliometric Analysis of Work on Human Gene Mapping, Samuel R. Reisher and Michael B. Albert, CHI Research/Computer Horizons, Inc., Haddon Heights, New Jersey (June 1987)

Medical Implications of Extensive Physical and Sequence Characterization of the Human Genome, Theodore Friedmann, Center for Molecular Genetics, School of Medicine, University of California-San Diego (November 1987)

Mapping the Human Genome: Some Implications, Jonathan Glover, New College, University of Oxford, Oxford, England (December 1987)

Mapping and Sequencing the Human Genome: Considerations from the History of Particle Accelerators, John L. Heilbron, University of California-Berkeley, and Daniel J. Kevles, California Institute of Technology (October 1987)

Mapping the Human Genome: Historical Background, Horace Freeland Judson (September 1987)

Long-term Implications of Mapping and Sequencing the Human Genome: Ethical and Philosophical Implications, Mark A. Lappé, College of Medicine, University of Illinois at Chicago, Chicago, Illinois (December 1987)

These are contractor reports used in preparing OTA's final assessment report. OTA makes them available for the use of readers desiring a more detailed or technical discussion of an issue than can normally be accommodated in our final report. As OTA contractor reports, they have not been reviewed or approved by the Technology Assessment Board. The findings and conclusions expressed are those of the authors and do not necessarily reflect the views of OTA, the assessment advisory panel or the Technology Assessment Board.

J952-7

This document can be ordered in either paper or microfiche format from:

National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Road
Springfield, Virginia 22161
Telephone: (703) 487-4650
Telex: 89-9405

TABLE OF CONTENTS

SECTION PAGE

I. INTRODUCTION 1

II. GROWTH OF THE RELEVANT SCIENTIFIC LITERATURE 2

III. ADDITIONAL INDICATORS OF THE RECENT PROGRESS IN MAPPING AND SEQUENCING THE HUMAN GENOME 10

- A. Number of Human Genes Identified and Mapped 10
- B. Number of Nucleotide Sequences Analyzed 11
- C. Largest Nucleotide Sequence Reported by Date 11

IV. MAJOR METHODS OF GENE MAPPING AND NUCLEOTIDE SEQUENCE DETERMINATION 19

- A. Methods of Gene Mapping 19
- B. Nucleotide Sequence Determination Strategy 19
- C. Sequence Determination 21

TABLE OF CONTENTS
(Continued)

APPENDICES

PART A. HUMAN GENE MAPPING 8 A-1

PART B. DATA FROM GENBANK FOR FIGURES 4 AND 5 A-5

PART C. COPY OF CHAPTER BY R. S. SPARKES "THE HUMAN GENE MAP" A-14

LIST OF TABLES

TABLE		PAGE
1	KEYWORD SEARCH OF MEDLINE DATABASE	3
2	FOLLOW-UP KEYWORD SEARCH OF MEDLINE DATABASE	6
3	NUMBER OF HUMAN GENES IDENTIFIED AND/OR MAPPED	13

LIST OF FIGURES

FIGURE		PAGE
1	LITERATURE GROWTH INDICATORS: HUMAN CHROMOSOME MAPPING	8
2	LITERATURE GROWTH INDICATORS: HUMAN BASE SEQUENCE	9
3	NUMBER OF HUMAN GENES IDENTIFIED AND/OR MAPPED	12
4	ESTIMATE OF PUBLICATION RATES FOR NUCLEOTIDE SEQUENCE DATA BY YEAR	14
5	GROWTH OF THE GENBANK NUCLEOTIDE SEQUENCE DATABASE	15
6	LARGEST PUBLISHED NUCLEOTIDE SEQUENCES	16

I. INTRODUCTION

The purpose of this project has been to assess the progress and the rate of advance of mapping and/or sequencing the human genome.

To achieve this purpose, three major approaches have been used:

1) identification and assessment of the growth of the relevant scientific literature by keyword searches;

2) development of specific indicators (e.g., number of genes mapped, number of nucleotide sequences analyzed, etc.); and

3) identification of important methods and strategies for mapping and/or sequencing the human genome.

These approaches and results are presented in the main body of this report.

A related project (OTA No. H3-4385.0) has been performed to provide a bibliography of the most relevant literature for the Office of Technology Assessment (OTA). This bibliography has been provided separate from this report.

II. GROWTH OF THE RELEVANT SCIENTIFIC LITERATURE

As an initial step in this project, a list of relevant topics and index terms was generated to provide a general overview of the growth of the relevant scientific literature over the past two decades. Table 1 provides the list of general search terms and the publication counts year-by-year, 1966-1986*. These publication counts were obtained by a search of the National Library of Medicine's MEDLINE database, as accessed through Dialog Information Services.

The use of publication counts as a measure of research activity is part of the field of bibliometrics. Bibliometric techniques are rapidly gaining acceptance on the basis of a growing body of research which demonstrates that counts of scientific papers and numbers of citations to these papers in later work are, respectively, legitimate indicators of research productivity and of quality or impact.

Minor limitations to bibliometric techniques have been acknowledged. Generally, these are overcome with adequate knowledge of the subject fields of research, adequate coverage of the output, careful handling of the data, and proper application of the techniques. In addition, certain adjustments must be made when dealing with foreign language publications, since the literature of science is dominated by English speaking researchers and there is an inherent bias

---

*The release of this report has been timed to permit inclusion of 1986 counts with reasonable confidence. It requires approximately six months after the end of the publication year (e.g., 1986), for "90%+" of the 1986 publications to have been entered into the database.

TABLE 1

KEYWORD SEARCH OF MEDLINE DATABASE*
(ACCESSED THROUGH DIALOG)

YEAR	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975
NUMBER OF RECORDS IN DATABASE	173,861	185,150	202,801	209,911	208,653	204,064	221,565	226,215	230,898	240,838
SEARCH TERMS										
CHROMOSOME										
Aberration	117	385	604	727	746	762	788	905	1,019	1,046
Abnormality	500	563	646	715	782	806	817	788	765	745
Banding	0	0	1	0	0	9	54	95	79	199
Deletion	17	24	15	26	23	24	34	32	46	139
Fragile Site								0	0	0
Mapping	155	225	255	368	283	329	450	541	926	897
Human	0	0	56	77	48	53	74	82	223	229
GENES										
Recessive	15	67	162	188	201	226	259	278	344	377
Reiterated	0	0	0	1	0	2	1	2	3	8
Structural	30	58	89	117	128	200	349	365	439	574
Synthetic	0	2	1	1	4	2	5	2	6	10
GENETIC										
Biochemical	341	498	506	516	421	415	354	299	300	347
Medical	327	366	388	400	321	312	282	299	241	268
Marker	2	7	2	3	5	6	5	12	7	19
HEREDITARY										
Disease	11	52	88	90	83	107	134	126	164	118
KARYOTYPE	403	643	769	852	882	909	1,034	1,097	1,215	991
NUCLEIC ACID	219	290	453	474	435	802	1,365	1,361	1,611	1,731
Hybridization	2	11	39	26	67	286	528	592	712	608
Repetitive Sequence	0	0	0	0	0	0	2	4	4	15
Sequence Homology	0	0	1	1	0	1	2	3	3	18
DNA										
Mitochondria	46	72	92	129	127	150	223	227	261	234
Restriction Enzyme	0	0	1	0	2	0	6	6	9	175
BASE SEQUENCE										
Human	0	1	0	0	1	22	43	55	98	

---

*Search covering years 1966 to 1985 performed February 1987, and for year 1986 performed June 1987.

TABLE 1
(Continued)

KEYWORD SEARCH OF MEDLINE DATABASE*
(ACCESSED THROUGH DIALOG)

YEAR	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986
NUMBER OF RECORDS IN DATABASE	249,285	246,829	264,944	262,915	265,027	275,194	283,722	295,673	300,169	291,926
SEARCH TERMS										
CHROMOSOME										
Aberration	815	644	701	629	656	699	720	697	858	811
Abnormality	673	699	798	696	766	800	905	933	1,021	981
Banding	221	257	359	330	397	413	381	441	532	648
Deletion	182	184	297	276	317	341	415	606	720	764
Fragile Site	1	0	4	5	32	43	47	60	59	93
Mapping	687	737	793	814	795	857	888	1,117	1,485	1,610
Human	185	267	252	287	220	359	342	498	727	807
GENES										
Recessive	268	256	274	276	251	353	317	340	344	337
Reiterated	15	15	11	14	22	19	15	31	14	58
Structural	685	828	905	1,307	1,494	1,774	2,066	2,838	2,600	2,932
Synthetic	21	25	41	70	69	45	69	134	174	221
GENETIC										
Biochemical	282	289	309	363	421	412	490	520	509	507
Medical	205	223	258	240	209	234	246	238	263	242
Marker	25	25	175	333	314	295	313	377	490	635
HEREDITARY										
Disease	149	136	127	148	112	125	122	126	171	168
KARYOTYPE	756	763	775	851	871	956	923	1,071	1,212	1,163
NUCLEIC ACID	1,671	1,621	1,835	2,035	2,273	2,468	2,802	3,375	3,714	4,022
Hybridization	645	657	805	831	960	1,129	1,291	1,780	1,871	2,186
Repetitive Sequence	13	21	25	120	298	359	404	428	564	523
Sequence Homology	26	26	39	43	39	53	65	102	355	815
DNA										
Mitochondria	190	184	250	239	218	240	277	295	281	153
Restriction Enzyme	389	482	728	917	1,203	1,490	1,688	2,449	2,243	2,275
BASE SEQUENCE										
Human	99	111	141	268	346	468	692	987	1,183	1,272

---

*Search covering years 1966 to 1985 performed February 1987, and for year 1986 performed June 1987.

against citations of foreign language publications. Finally, citation data must be carefully normalized for subject area and citation time span. For example, papers in such basic areas as biochemistry are much more highly cited than papers in very clinical fields (F. Narin, "Bibliometric Techniques in Evaluation of Research Programs". Science and Public Policy, 14, No. 2:99-106, 1987).

Following a review of the retrieved data by CHI Research and OTA, a refined search list was developed and used as the basis for a second search of MEDLINE. Table 2 provides the list of selected search terms and the year-by-year publication counts. This second search provides the basis for the selected bibliography prepared for OTA under Contract No. H3-4385.0.

While it is realized that many additional or alternative searches could be constructed, it should be realized that the primary purpose of this search was to provide indicators of growth rather than to develop specific bibliographies.

Figures 1 and 2 graphically display the dramatic acceleration of the growth of the scientific literature in two of these topics central to the mapping and sequencing of the human genome.

In conclusion, this bibliometric approach clearly demonstrates the very rapid growth in the scientific literature related to mapping and sequencing the human genome. Moreover, the high acceleration in this growth over the very recent years is evident.

TABLE 2

FOLLOW-UP KEYWORD SEARCH OF MEDLINE DATABASE*

YEAR	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975
NUMBER OF RECORDS IN DATABASE	173,861	188,180	202,801	209,911	208,663	204,064	221,565	226,218	230,898	240,838
GENE/GENES/GENETIC	3,488	4,826	6,129	6,868	6,929	7,038	8,477	8,941	10,397	10,903
Human	2,064	2,763	3,454	3,658	3,903	3,875	4,608	4,910	5,652	5,663
GENE/GENES/GENETIC MARKER	17	27	28	40	41	41	49	62	83	322
Human	9	6	13	15	13	16	20	25	53	133
GENE/GENES/GENETIC MAP	116	189	244	332	261	297	404	495	841	884
Human	15	27	58	63	46	43	63	76	189	209
CHROMOSOME	1,336	1,725	2,100	2,513	2,611	2,706	2,921	3,127	3,680	3,892
Human	709	907	1,177	1,351	1,431	1,482	1,543	1,664	1,843	1,965
DNA-SEQUENCE	20	24	45	53	61	126	279	331	470	665
Human	4	4	3	5	9	16	38	24	79	113
NUCLEIC ACID SEQUENCE	4	1	16	15	15	87	227	304	379	470
Human	0	0	0	0	2	8	101	172	214	250
NUCLEIC ACID	219	290	453	474	435	802	1,365	1,361	1,611	1,731
Automated	0	0	0	0	1	1		0	1	4
Analysis	0	0	0	0	1	0	1	0	0	3
Sequence								0	0	0
Sythensis	0							0	0	0
Method	4	16	47	65	37	71	112	95	130	201
Sequence	0	0	3	2	0	3	17	15	23	52
Technique	2	1	3	2	1	2	5	5	4	83
CLONE/CLONING	68	133	223	242	289	334	495	638	720	704
Fragment	0	0	0	0	2	1	3	16	13	28
Large	0	0	0	0	0	0	0	0	0	2
Size	0	0	0	0	0	0	0	0	0	2
RESTRICTION FRAGMENT LENGTH POLYMORPHISM (RFLP)										
Human										
DNA	2,120	2,584	3,163	3,584	3,839	4,190	4,920	5,306	6,038	6,.16
Sequence	20	24	45	53	61	126	280	335	472	687
Human	4	4	3	5	9	16	38	44	81	116
DNA AND AUTOMATED	1	0	2	0	2	7	7	3	7	13
Analysis	1	0	2	0	1	5	4	3	6	11
Sequence	0	0	0	0	1	0	1	4	2	24
Synthesis	0	0	0	0	0	0	0	0	0	4
DNA AND METHOD	61	108	263	370	323	323	422	365	467	651
Sequence	0	2	4	6	3	8	23	23	41	101
DNA AND TECHNIQUE	4	8	11	11	11	8	24	29	21	261
DNA OR NUCLEIC ACID	2,205	2,691	3,300	3,724	4,000	4,676	5,303	5,693	6,489	7,253
Sequence	23	25	48	54	68	171	346	418	565	788
Human	4	4	3	5	11	20	44	51	85	124
DNA OR NUCLEIC AND AUTOMATED	7	5	7	10	11	27	22	11	25	41
Analysis	5	1	4	3	3	14	9	8	13	23
Sequence	0	0	0	0	0	0	1	3	2	26
Synthesis	0	0	0	0	0	0	0	0	0	5
Method	4	2	3	5	6	14	6	3	15	19
Sequence	0	0	0	0	1	0	1	3	2	22
Technique	0	0	1	0	0	0	0	0	0	8
NUCLEOTIDE	1,284	1,560	2,018	2,364	2,555	2,365	2,116	1,803	1,944	2,208
Sequence	28	38	49	77	70	119	197	209	242	348
Human	0	5	3	2	3	8	13	22	22	39
Automated	1	1	1	4	1	5	1	2	1	1
Analysis	1	0	1	3	1	5	1	1	0	1
Sequence	0	0	0	0	0	0	0	0	0	1
Synthesis	0	0	0	0	0		0	0	0	0
Method	17	38	164	363	336	167	165	132	173	226
Sequence	0	2	4	20	9	13	18	16	20	49
Technique	2	2	3	3	6	2	3	2	7	62

---

*Search covering years 1966 to 1985 performed February 1987, and for year 1985 performed June 1987.

TABLE 2
(Continued)

FOLLOW-UP KEYWORD SEARCH OF MEDLINE DATABASE*

YEAR	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986
NUMBER OF RECORDS IN DATABASE	234,452	249,285	246,829	264,944	262,940	268,027	275,211	286,786	296,781	301,446	291,926
GENE/GENES/GENETIC	10,706	11,366	12,069	13,809	14,342	16,621	17,597	19,134	20,600	21,862	22,291
Human	5,753	5,961	6,236	6,858	7,018	7,557	8,562	9,346	10,280	11,349	11,668
GENE/GENES/GENETIC MARKER	369	350	414	551	640	689	761	908	1,010	1,164	1,378
Human	167	187	193	268	365	373	413	523	632	790	948
GENE/GENES/GENETIC MAP	618	777	987	1,159	1,220	1,325	1,489	1,690	2,032	2,336	2,357
Human	276	184	285	292	329	320	468	527	697	917	1,014
CHROMOSOME	3,483	3,507	3,518	3,830	3,756	3,940	4,147	4,395	4,889	5,441	5,428
Human	1,847	1,876	1,896	2,040	1,984	2,088	2,246	2,404	2,738	3,179	3,317
DNA-SEQUENCE	648	771	986	1,332	1,687	2,058	2,600	3,208	4,172	4,677	4,902
Human	91	109	138	186	316	437	603	908	1,181	1,476	1,718
NUCLEIC ACID SEQUENCE	466	533	597	794	873	1,120	1,308	1,562	2,014	2,133	2,556
Human	187	202	190	147	143	228	273	413	577	716	944
NUCLEIC ACID	1,626	1,671	1,621	1,836	1,847	2,071	2,257	2,600	3,170	3,519	4,022
Automated	4	1	1	6	0	1	3	5	8	5	9
Analysis	2	1	1	5	0	0	1	4	3	1	7
Sequence	2	0	0	0	0	0	2	4	7	2	6
Sythenesis	2	0	0	0	0	0	1	0	1	3	1
Method	177	173	200	240	227	204	198	284	364	409	499
Sequence	30	41	64	98	114	94	100	124	177	176	238
Technique	80	70	86	90	103	122	132	146	192	240	270
CLONE/CLONING	606	730	873	1,083	1,498	2,162	2,325	3,022	3,728	4,001	4,174
Fragment	18	46	69	118	270	362	444	517	680	707	748
Large	2	3	5	43	96	159	148	234	261	309	260
Size	1	4	2	37	108	131	124	185	284	288	287
RESTRICTION FRAGMENT LENGTH POLYMORPHISM (RFLP)					2	1	11	34	74	192	166
Human				0	2	1	9	31	68	162	152
DNA	5,983	6,308	6,537	7,246	7,800	8,106	9,213	10,075	11,845	12,747	13,164
Sequence	636	784	998	1,349	1,698	2,068	2,609	3,217	4,185	4,485	4,902
Human	91	109	137	191	320	443	605	909	1,185	1,479	1,718
DNA AND AUTOMATED	13	15	9	25	19	13	25	35	49	44	38
Analysis	9	10	5	20	16	9	15	27	34	34	26
Sequence	13	14	13	22	13	17	18	21	32	22	19
Synthesis	1	2	2	4	0	4	3	0	7	9	8
DNA AND METHOD	616	608	718	819	970	874	902	1,128	1,298	1,377	724
Sequence	67	78	138	181	249	224	282	300	389	360	244
DNA AND TECHNIQUE	237	246	298	327	374	400	437	556	688	723	717
DNA OR NUCLEIC ACID	6,852	6,870	7,170	7,740	8,308	8,603	9,709	10,616	12,467	13,539	14,037
Sequence	838	926	1,173	1,526	1,878	2,274	2,809	3,472	4,477	4,863	5,520
Human	102	120	182	197	341	477	634	989	1,248	1,606	1,929
DNA OR NUCLEIC AND AUTOMATED	46	47	38	66	46	49	56	75	99	78	75
Analysis	27	29	26	49	32	27	33	64	67	57	61
Sequence	16	16	16	14	8	8	8	17	17	20	26
Synthesis	3	1	1	5	0	0	3	2	8	5	8
Method	24	24	16	37	22	22	29	31	48	38	34
Sequence	13	13	12	23	12	13	16	13	21	11	4
Technique	9	6	5	13	7	7	7	12	11	11	11
NUCLEOTIDE	1,973	2,104	2,000	2,088	2,482	2,364	2,686	3,041	3,614	3,871	3,932
Sequence	296	317	386	488	616	718	927	1,178	1,810	1,722	1,918
Human	28	38	[illegible]	49	116	143	208	302	402	536	570
Automated	4	1	6	3	3	4	6	10	9	13	17
Analysis	3	1	3	1	2	1	5	9	6	9	10
Sequence	2	0	3	1	0	3	5	6	8	11	13
Synthesis	1	0	0	1	0	0	0	0	4	1	4
Method	190	214	220	222	301	232	224	263	336	334	319
Sequence	39	38	67	66	91	88	87	92	128	132	141
Technique	73	68	66	73	97	96	98	118	121	147	128

---

*Search covering years 1966 to 1985 performed February 1987, and for year 1985 performed June 1987.

FIGURE 1

LITERATURE GROWTH INDICATORS

HUMAN CHROMOSOME MAPPING

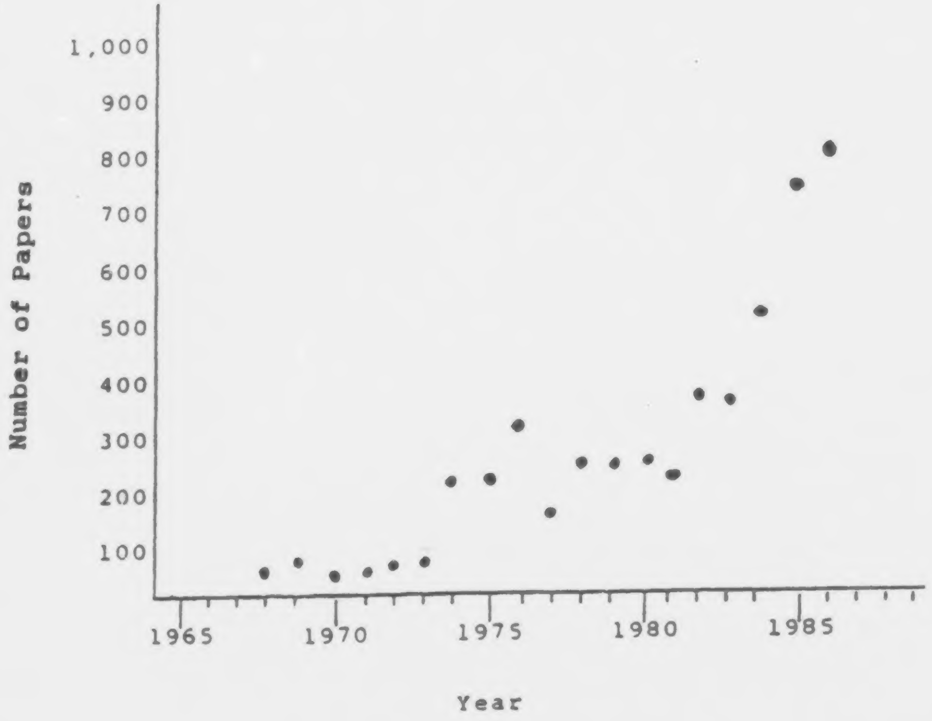
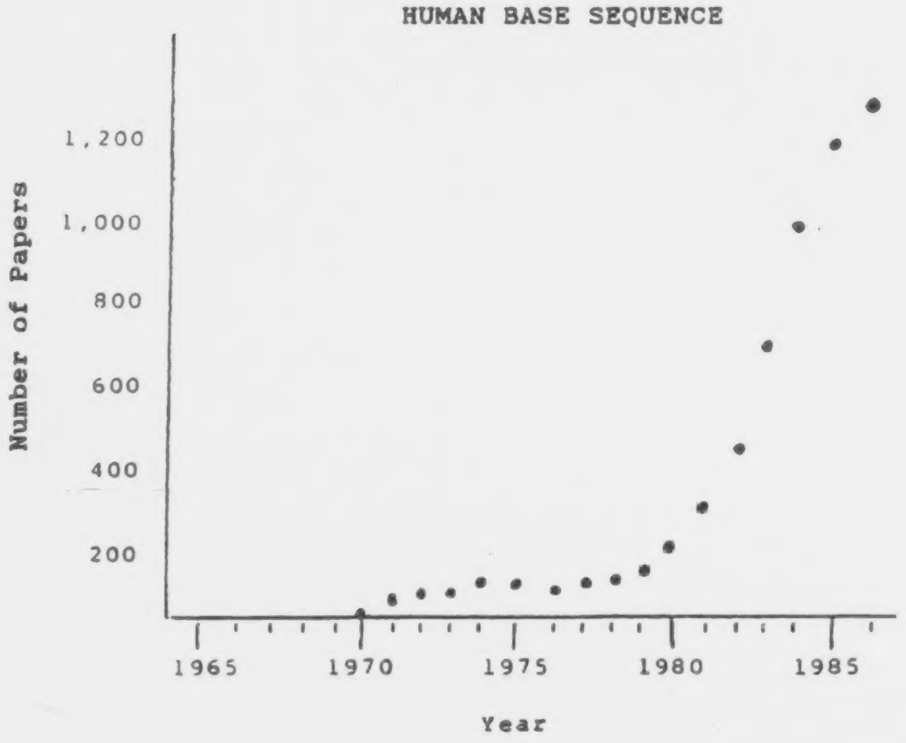


FIGURE 2

LITERATURE GROWTH INDICATORS



III. ADDITIONAL INDICATORS OF THE RECENT PROGRESS IN MAPPING AND SEQUENCING THE HUMAN GENOME

While the literature indicators provide good measures of overall research activity in the areas of interest to this project, it was clear from the outset that several specific indicators of the actual progress in mapping and sequencing the human genome could be developed.

The most obvious specific indicators are the number of genes identified, the number of genes mapped, and the number of sequences (i.e., nucleotide base pairs) analyzed. An additional important indicator is the largest contiguous (non-trivial) nucleotide sequence obtained. Attempts were also made to identify the largest reported cloned fragment to date. Each of the indicators is described below.

A. Number of Human Genes Identified and Mapped

While the number of human genes identified, and number of human genes mapped, seem at first glance, to be relatively unambiguous indicators of progress, certain considerations of the ambiguity in these concepts should be made. First, as our understanding of the structure and function of the cell has evolved, so has our concept of the "gene". Thus, from "one gene-one enzyme" hypothesis to "one gene-one polypeptide"; from transcription of one contiguous nucleotide sequence ("gene") to primary transcript processing of "split genes" i.e., removal of intervening sequences ("introns") and splicing of "exons" to form an RNA; from hybridization of complementary DNA (cDNA) formed from known RNAs to hybridization of cloned "anonymous" cDNA. No doubt our definition of "gene" will continue to evolve, as our understanding of the genome evolves.

Just as our concept of "gene" has evolved, so has our definition of "mapping". Early on, mapping was the identification of the chromosome on which a given marker was carried. Now, "regional" assignments and more localized assignments are possible with the continuing refinements in methodology.

With these considerations in mind, Figure 3 presents a composite of the number of human genes identified and mapped through time. Table 3 provides the data for Figure 3 (See also Appendix, Part A, "Human Gene Mapping 8").

B. Number of Nucleotide Sequences Analyzed

A very clear measure of progress in sequencing the human genome is obtainable by plotting the number of human nucleotide base pairs analyzed as a function of time. While this data is obtainable by an analysis of the GenBank Nucleotide Sequence Data Base, we have not obtained this data. Rather, we present data provided by Dr. Wayne Rindone of GenBank, showing growth of total nucleotide sequence data (including non-human), and the human data just for the most recent years.

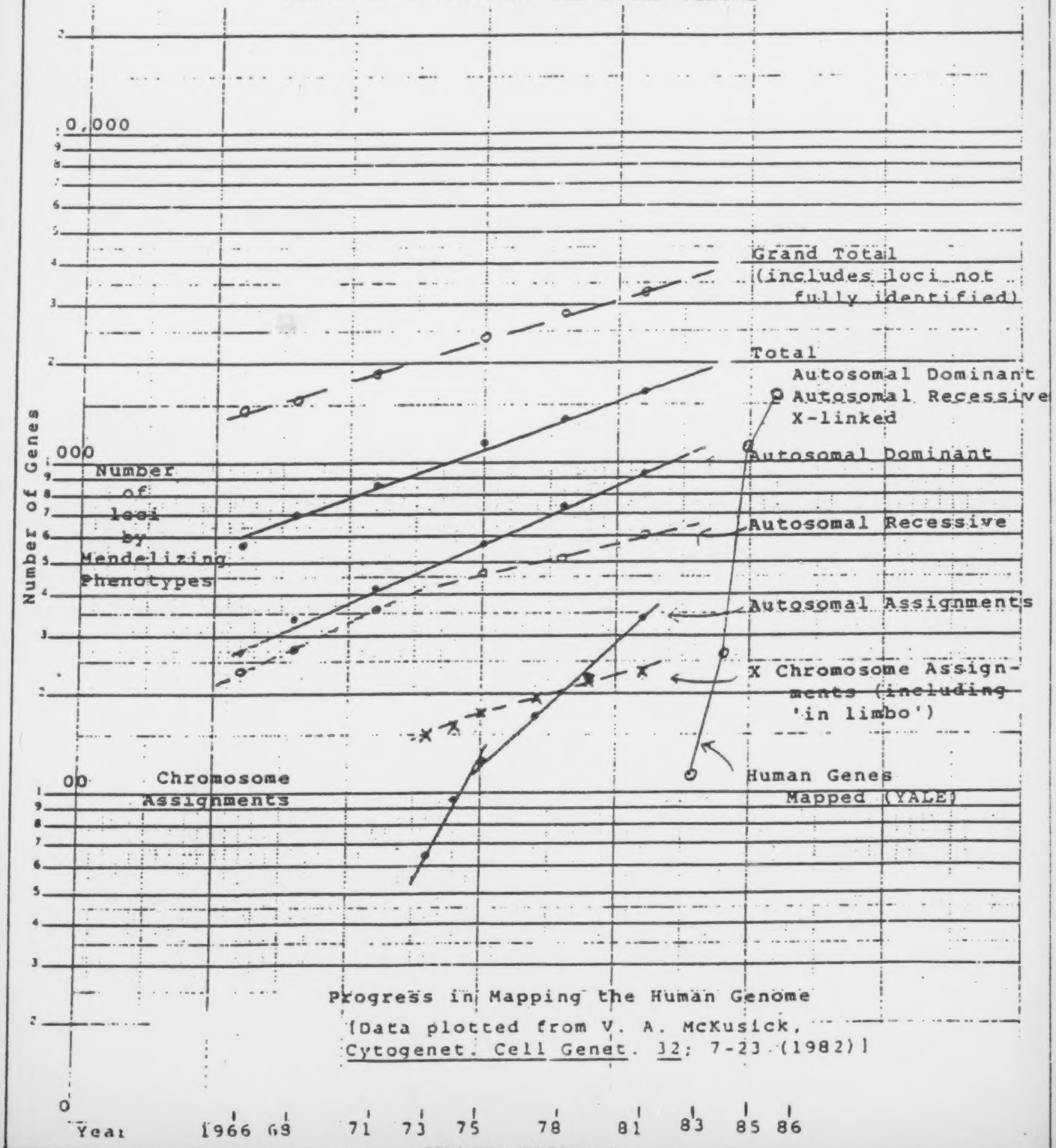
These data are provided in the Appendix, Part B and plotted as Figures 4 and 5.

C. Largest Nucleotide Sequence Reported by Date

Perhaps the single most relevant measure of our collective ability to arrive at the nucleotide sequence of a target DNA is the largest nucleotide sequence reported (Figure 6). This indicator represents the integration of each of the steps required to identify, amplify, isolate, purify, analyze, and synthesize the fragmented sequence data into the

FIGURE 3

PROGRESS IN MAPPING THE HUMAN GENOME



Progress in Mapping the Human Genome
[Data plotted from V. A. McKusick, Cytogenet. Cell Genet. 32: 7-23 (1982)]

TABLE 3

NUMBER OF HUMAN GENES IDENTIFIED AND/OR MAPPED

A. **Table III. Number of loci identified mainly by mendelizing phenotypes (McKusick, 1978, with additions)[1]**

Phenotype	Verschuer (1958)	Mendelian Inheritance in Man					
		1966	1968	1971	1975	1978	1981
Autosomal dominant	285	269 (+568)	344 (+449)	415 (+528)	583 (+635)	736 (+753)	918 (+883)
Autosomal recessive	89	237 (+294)	280 (+349)	365 (+418)	466 (+481)	521 (+596)	601 (+692)
X-linked	38	68 (+51)	68 (+55)	86 (+64)	93 (+78)	107 (+98)	116 (+123)
Total	412	574 (+913)	692 (+853)	866 (+1,010)	1,142 (+1,194)	1,364 (+1,447)	1,635 (+1,698)
Grand total	412	1,487	1,545	1,876	2,336	2,811	3,333

[1] McKusick, V.A.: Mendelian Inheritance in Man, 5th Ed., 1978. The first edition was published in 1966, the second in 1968, the third in 1971, the fourth in 1975, and the fifth in 1978. The numbers given under 1981 are from unpublished additions to this catalogue, October 26, 1981. Numbers in parentheses refer to loci not yet fully identified or confirmed.

B. Table I. Number of autosomal loci assigned at each of the six human gene mapping workshops

Conference	Number of autosomal assignments				X chromosome assignments	
	Confirmed	Provisional	Inconsistent[2]	Total	Confirmed	In limbo
New Haven (1973)	31	28	5	64	88	67
Rotterdam (1974)	48	32	6	86	91	70
Baltimore (1975)	72	46	7	125	95	80
Winnipeg (1977)	83	82	11	176	102	96
Edinburgh (1979)	123	87	20	230	112	101
Oslo (1981)	180	120	45	345	116	118

[1] Numbers for first four conferences provided by Donald and Hamerton (1978).
[2] Including assignments in limbo.

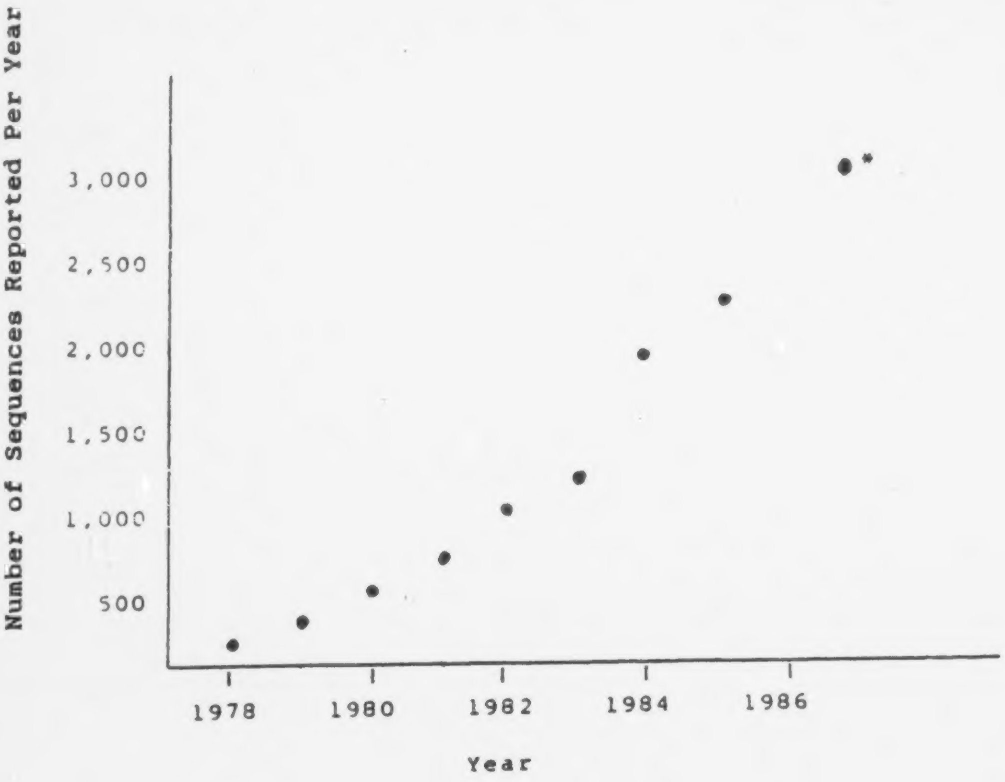
C. Number of Genes Mapped - Yale Human Gene Mapping Library

Year	<1984	1984	1985	1986
Reported Mapped	113	158	801	541
Cum Total	113	271	1,072	1,613

See also "Human Gene Mapping 8" in Appendix.

FIGURE 4

ESTIMATE OF PUBLICATION RATES FOR NUCLEOTIDE SEQUENCE DATA BY YEAR

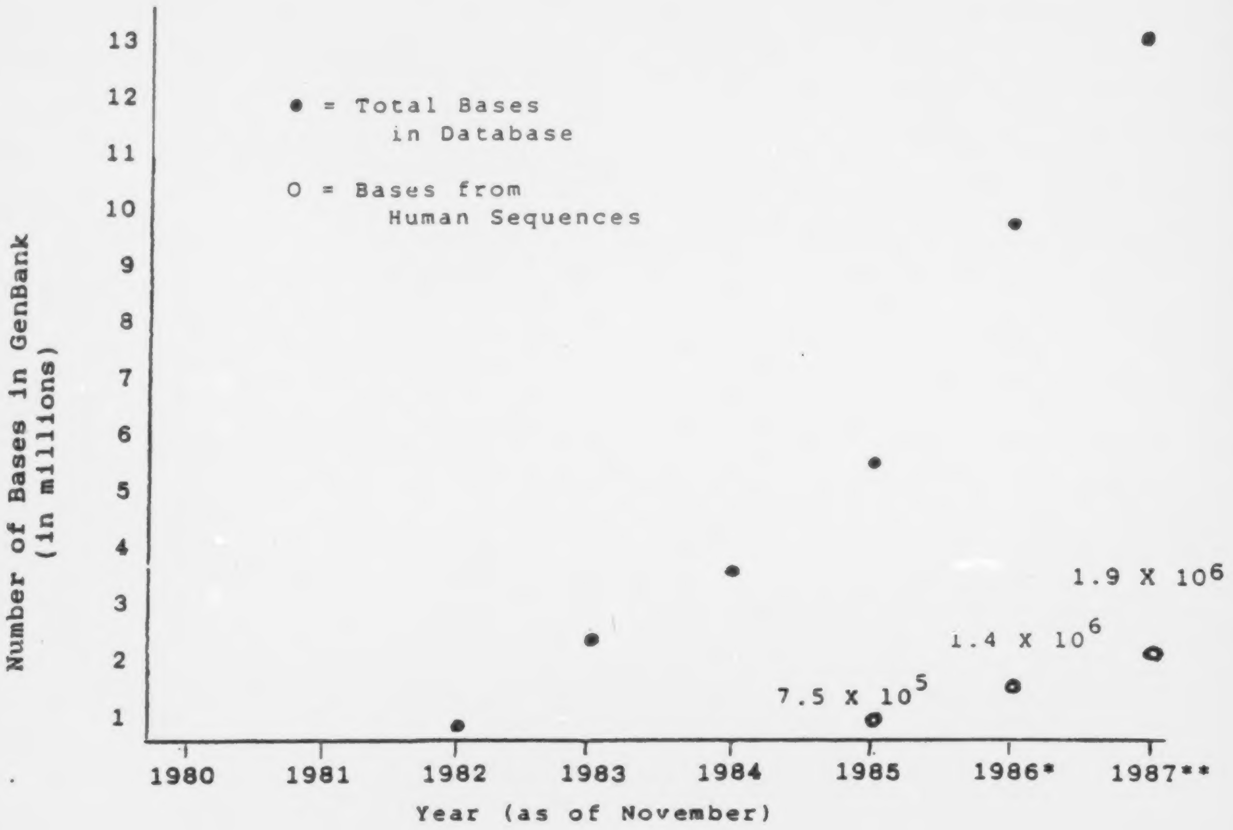


*May 1987 data (GenBank release 50.0)

(Data provided by Dr. Wayne P. Rindone, GenBank).

FIGURE 5

GROWTH OF THE GENBANK NUCLEOTIDE SEQUENCE DATABASE



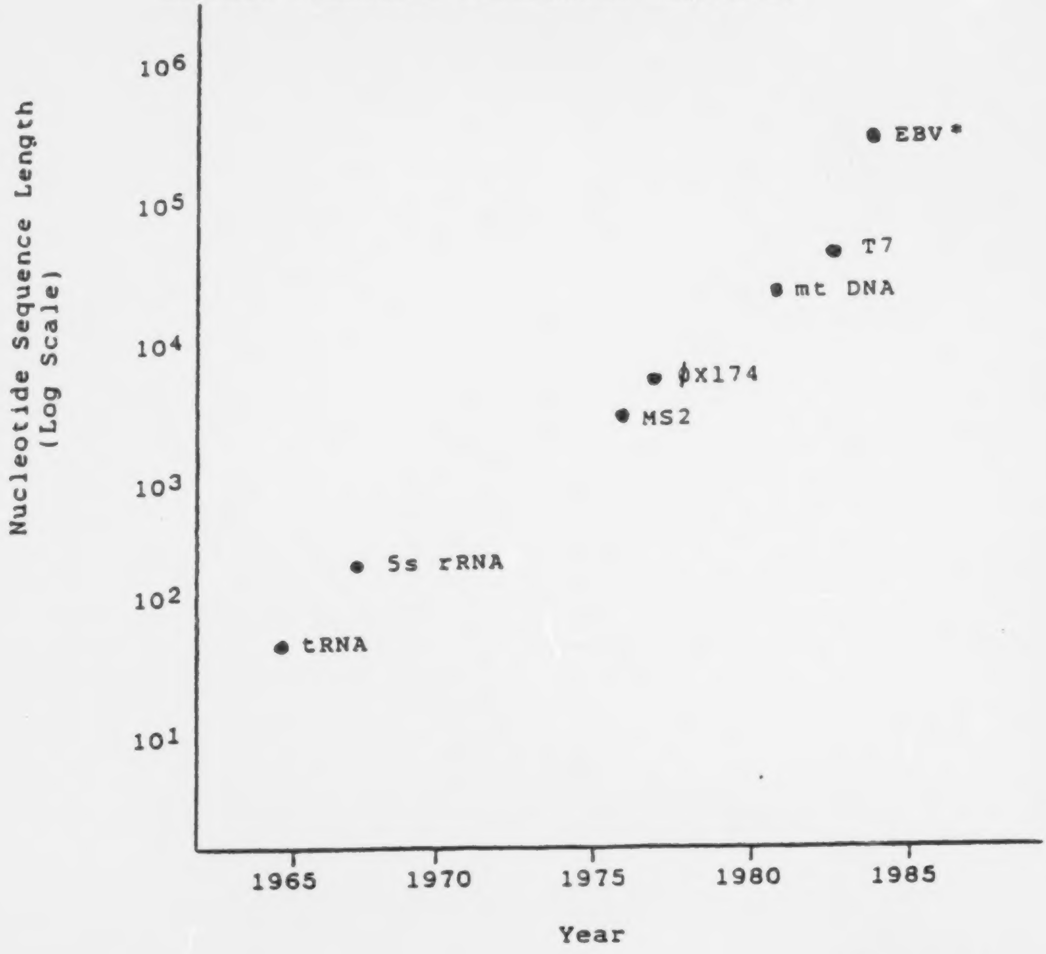
<N> The human sequence data has not been tabulated separately for earlier years, but counts could be extracted, if needed.

*Part of the increase for 1986 is due to the inclusion of unannotated sequences (1.2 x 10^6 bases), as well as from the growing input from the EMBL Nucleotide Sequence Data Library.

**May 1987 data (GenBank release 50.0)
(Data provided by Dr. Wayne P. Rindone, GenBank).

FIGURE 6

LARGEST PUBLISHED NUCLEOTIDE SEQUENCES



Date	Sequence	Bases or Base Pairs
1965	Alanyl tRNA	77
1968	5S Ribosomal RNA (5S-rRNA)	120
1976	Bacteriophage MS2 (RNA)	3,569
1977	Bacteriophage ØX174	5,375
1981	Human Mitochondrial DNA (mt DNA)	16,569
1983	Bacteriophage T7	39,936
1984	Epstein-Barr Virus (EBV)*	172,282

*Largest published sequence as of June 1987 (Source: Dr. Wayne P. Rindone, GenBank).

complete contiguous (unambiguous) nucleotide sequence of the target DNA. While the final published sequence represents the cumulative efforts through time of the research team (including a learning curve, concomitant improvements in methods, and perhaps, improvements in strategy), the results (i.e., the published nucleotide sequence) represents an end point for which resources (time, money, personnel) can be calculated. Thus, by analogy, resources required to elucidate the nucleotide sequence of alternative DNA targets of similar or increased complexity (size) may be estimated, without recourse to additional anticipated, but not yet realized, improvements in methods, strategies, etc. That is, experience can serve as a reliable base for future endeavors.

While a plot of the largest nucleotide sequence reported through time demonstrates accomplishments to date, implicit in the plot is the rate of advance of improvements in sequencing methods and strategies. That is, the plot provides not only a record of the achievements to date, but also a reasonable estimate of what is feasible to accomplish in the near future. Extrapolation to more distant times is questionable.

What this implies is that, if one's goal is to achieve the complete sequence of circa 3×10^9 base pairs of the human genome, it might be prudent to set intermediate goals of contiguous sequence data at perhaps one order of magnitude above (i.e., 10X) most recent accomplishments. This approach suggests a more conservative strategy than that advocated by others. However, it promises to accomplish within its purview other intermediate targets (e.g., E. coli, yeast, nematode, Drosophila, etc.) up to one small human chromosome, and may encompass a wider constituency than would a forthright attack upon the entire human genome. It is generally conceded that improvements in methodology promise orders of magnitude improvements in cost-effectiveness, and that any apparent delay in making the initial assault on the human genome will be more than compensated for in cost and time efficiencies achievable by near-term advances in technology.

IV. MAJOR METHODS OF GENE MAPPING AND NUCLEOTIDE SEQUENCE DETERMINATION*

There are a large number of particular methods for gene mapping and nucleotide sequence determination. There are modifications and refinements of each method for various applications to different cell types, different target genes, different degrees of resolution, and different size ranges of target nucleic acid molecules. Rather than identify the specific methods, we here list some of the major approaches. Other than for a few classic references, we omit specific references.

A. Methods of Gene Mapping

Rather than replicate the effort of others and attempt a summary of these methods which would be much less complete and less concisely presented than already available, we here provide a table extracted from the work of Robert S. Sparkes ("The Human Gene Map"). A copy of this paper is appended to this report (Appendix, Part C) for easy reference. This paper provides a brief history, concise descriptions of the various methods, and discussions of the clinical relevance of the human gene map.

B. Nucleotide Sequence Determination Strategy

Most common strategies to determine the nucleotide sequence of a target DNA genome or large segment (say, more than a few hundred nucleotides) of DNA involve the following steps:

---

*To a certain extent, CHI Research considers this beyond the purview of the present contract.

a) isolate the target DNA;

b) develop a restriction map of the target;

c) develop a "library" of cloned fragments which cover the entire target;

d) sequence the appropriate cloned fragments; and

e) assemble the sequence data from the individual fragments into the complete target sequence.

At the present time, the following may be taken as a measure of our collective capability to apply these strategies successfully to large targets.

1) The complete nucleotide sequence (172,282 base pairs) of the Epstein-Barr virus genome has been obtained (Baer et al., Nature, 310: 207-211 (1984); and

2) A complete restriction map of the nematode Caenorhabditis elegans (genome; ~8 X 10^7 base pairs) is being attempted (Coulson et al., Proc. Nat. Acad. Sci., 83: 7821-7825 (1986)).

The first point represents a milestone already passed, while the second point represents our "reach" if not quite yet our "grasp".

The particular relevance of the second point to the prospects for mapping and sequencing the human genome is that the nematode genome is in the size range of the smaller human chromosomes (10^7 to 10^8 base pairs). In this regard it should be mentioned that large, but not complete, human chromosome--specific libraries have already been created (see announcement by American Type Culture Collection of the "Repository of Human DNA Probes and Libraries"--National Laboratory Gene Library Project in Science, August 22, 1986.)*

C. Sequence Determination

The classic references for DNA sequencing are:

1) A.M. Maxam, and W. Gilbert, "A New Method for Sequencing DNA". Proc. Nat. Acad. Sci., 74: 560 (1977); and

2) F. Sanger, S. Nicklen, and A. R. Coulson, "DNA Sequencing with Chain-Terminating Inhibitors". Proc. Nat. Acad. Sci., 74: 5463 (1977).

There have been subsequent refinements and improvements to these basic methods which have increased the number of contiguous nucleotides deductible (per run) from about 150-200 up to about 300. Each of these methods employ radio-labels followed by autoradiography.

More recently, Smith, et. al. reported the use of fluorescent labels as part of an automated DNA sequence analysis system (L. M. Smith, J. Z. Sanders, R. J. Kaiser, P.

---

*See also Nature, 323: 660 (1986) reporting the relatively high contamination of the library with hamster DNA.

Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. H. Kent, and L. E. Hood, "Fluorescence Detection in Automated DNA Sequence Analysis". Nature, 321: 674-679 (1986). Sequence information of the order of 200 to 400 bases (per run) has been obtained by the fluorescence method.

An alternative automated DNA sequencing machine has been reported by Wada (A. Wada, "Automatic DNA Sequencing" Nature, 307: 193 (1984). As part of a proposed DNA sequencing system, "Wada estimates that it should be possible to sequence one million base pairs per day" Nature, 322: 397 (1986).

While 10^6 base pairs per day is impressive, it should be considered that the ~3 X 10^9 base pairs of the human genome would still require ~3,000 days or about eight years to sequence. Moreover, the capacity to obtain sequence data for millions or billions of base pairs does not guarantee the correct ordering of the subsequence data (say of cloned restriction fragments) unambiguously into that of the target genome. Indeed, it is to be expected that there is some degree of heterogeneity in the sequence data due to potential allelic variation, spontaneous mutation, infidelity of replication, and random or systematic errors in the sequencing process. Thus, at a minimum, one would anticipate at least a two to ten-fold redundancy in the sequencing process to identify and evaluate endogenous variability, as well as to identify and minimize any systematic errors.

Since humans, like snowflakes, are non-identical (and indeed perhaps the DNA in any one of our cells is different from that in any other cell) the definition of "the" sequence of the human genome is problematic.

APPENDIX

PART A

HUMAN GENE MAPPING 8

Human Gene Mapping 8

Helsinki Conference (1985)
Eighth International Workshop on Human Gene Mapping

Sponsored by
March of Dimes Birth Defects Foundation, USA
Ministry of Education, Finland
Pension-Varma Insurance Co., Finland
National Science Foundation, USA
and the organizations listed on page 7

at the University of Helsinki, Finland
August 4–10, 1985

Executive Committee:
Albert de la Chapelle, M.D.
Robert S. Sparkes, M.D.
Kåre Berg, M.D.

Organizer:
Albert de la Chapelle, M.D.

Assistant Organizer:
Kalle Simola, M.D.

Editor:
Albert de la Chapelle, M.D.

Editorial Secretary:
Vivian Schultz

Cytogenetics and Cell Genetics:
Editor-in-Chief: Harold P. Klinger, M.D., Ph.D.
Assistant Editor: Margot Kaelbling, Ph.D.
Editorial Secretary: Barbara Scuola

Birth Defects: Original Article Series
March of Dimes
Birth Defects Foundation

Vol. 21, No. 4, 1985

Cytogenetics and Cell Genetics
S. Karger
Basel · München · Paris · London · New York · New Delhi · Singapore · Tokyo · Sydney

Vol. 40, Nos. 1–4, 1985

JUN 1 1986
BIOMEDICAL LIBRARY

The 1985 human gene map and human gene mapping in 1985

A. de la Chapelle

The human gene map in 1985

When the participants of the Eighth Human Gene Mapping Workshop left Helsinki on August 10, 1985 the total number of human genes and markers mapped at the end of HGM8 was 1479. Included are 559 cloned arbitrary DNA segments that have been assigned to a chromosome or part thereof. A total of 89 fragile sites are also included in the figure. Not included are 85 recurrent breakpoints occurring in association with specific types of neoplasia.

Table I provides a summary of how the numbers of assignments have developed over the last four years, between HGM6 and HGM8. The corresponding numbers are shown by chromosome in Table II. A few comments are necessary. At HGM6 only a small number of genes assigned to the X chromosome were included in the total figure of mapped genes (319) whereas at HGM7 the figure (620) included many but not all X chromosomal genes without regional assignment. The HGM8 figure of 831 mapped genes includes many (but not all) known X chromosomal genes without regional localization. In Table II the figures for the X chromosome include only regionally assigned genes, but all cloned DNA segments assigned to the X. Other sources of apparent discrepancies between the figures in Tables I and II stem from variations in the listing and reporting of fragile sites, antigens identified by monoclonal antibodies, protein spots in 2-dimensional gels, gene clusters, and unassigned linkage groups. For explanations regarding unmapped but cloned genes, gene clusters, and DNA segments

Table I. Numbers of mapped genes and markers, cloned DNA segments, fragile sites, cloned genes, and neoplasia-associated breakpoints at HGM6, HGM7 and HGM8. For explanations please refer to the text.

	HGM6 1981	HGM7 1983	HGM8 1985
Genes and markers including cloned genes	319	620	831
Cloned arbitrary DNA segments	35	215	559
Fragile sites	0	20	89
Total genes, markers, cloned DNA segments, and fragile sites	354	855	1479
Cloned genes	16	104	249
Total cloned DNA segments (genes plus arbitrary segments)	51	319	808
Neoplasia-associated breakpoints	12	44	85

Table II. Assignment by chromosome of genes and markers and of cloned DNA segments. For explanations please refer to the text.

Chromosome	Genes and markers[1,2]			Cloned DNA segments[2]		
	HGM6	HGM7	HGM8	HGM6	HGM7	HGM8
1	37	48	86	2	14	31
2	17	18	50	4	7	25
3	6	13	26	1	7	17
4	11	21	33	0	18	49
5	7	12	28	1	6	11
6	33	47	76	3	27	31
7	14	28	45	3	11	23
8	5	11	21	2	5	8
9	13	23	35	0	9	11
10	10	16	36	0	3	8
11	20	29	67	14	22	41
12	13	26	44	1	12	22
13	3	7	17	0	9	35
14	12	22	33	3	8	21
15	12	20	26	0	6	13
16	18	25	31	4	6	9
17	8	17	29	4	13	18
18	2	3	15	0	4	9
19	11	23	38	0	7	22
20	4	7	14	3	6	9
21	9	12	14	0	21	69
22	9	11	21	1	10	19
X	9[3]	25[3]	54[3]	10	93	214
XY	0	0	1	0	11	27
Y	3	7	7	2	13	29
Unassigned	33/319	37/508	22/869	10	28	18

[1] Including fragile sites; [2] Including cloned genes; [3] Including only regionally assigned genes and markers

mapped to several chromosomes, please refer to the report of the Committee on Human Gene Mapping by Recombinant DNA Techniques (HGM8). These variations in reporting are apparently also responsible for those considerable differences in the figures that have occurred in the literature on human gene mapping.

Of the total 920 mapped genes, markers and fragile sites, 404 were considered confirmed, 476 provisional, 19 tentative, and 5 inconsistent. For definitions of these designations, see below. In the case of 16 genes belonging to unassigned linkage groups these designations were not applicable. As in the past, the majority of assignments that were previously listed as provisional have been confirmed. The high proportion of provisional assignments is in part due to many recently described fragile sites which have been studied in one laboratory only. The numbers of assignments listed as tentative and inconsistent remain low; indeed certain committees did not use these categories.

APPENDIX

PART B

DATA FROM GENBANK FOR FIGURES 4 AND 5

Publication Rate of Sequence Data
and
Rate of Appearance of the Data in the GenBank Database

Updated August 1986

This document was originally prepared in March 1986, using data gathered from the March 1986 release of the GenBank database. It has been updated to also include data from the August 1986 GenBank release.

The routine task of scanning the journals that publish sequence data is split between GenBank and EMBL. In an independent audit of that process, an expanded set of journals was systematically sampled for the years 1978 through 1985, providing an overall picture of the publication of sequence data.

First, the June 1985 issues of the expanded list were scanned to see which journals published a "significant" number of sequence papers, and the list pared to 24 journals that can be expected to publish at least 10 papers per year. For each of the 24 and each of the years 1978-1985 the June issue(s) were scanned for papers with figures presenting what appears to be a nucleic acid sequence. Our experience indicates that 95% of such papers contain original data. For each journal the number of expected sequences per annum is computed from the June sample. For a few journals not available to us an estimate was based on titles from CURRENT CONTENTS.

From several checks on the method we conclude that a) the journals not included in the main scan contribute less than 5% of all sequence data, and b) that the publication rates estimated from the June issues are probably about 15% high; more sequence papers seem to be published in the summer.

For CSHSQB (Cold Spring Harbor Symposia) the first 1/4 of the symposia proceedings were examined and the results multiplied by 4. Where there are no data available to us, projections from other years are given in parentheses.

Results, numbers of sequences.

Journal	1985	1984	1983	1982	1981	1980	1979	1978	How
BBRC	36	0	0	0	0	0	0	0	Lib
Biochem-USA	39	13	13	13	39	0	13	0	Lib
Bioorg Khim	48	12	0	36	0	0	0	0	Lib
Cell	150	252	120	108	108	120	24	24	Lib
CSHSQB	(20)	20	8	12	0	36	16	40	Lib
Curr Gen	48	72	60	(50)	(40)	(30)	(20)	(10)	CC
DNA	43	48	48	48	0	0	0	0	CC
EMBO J	180	108	84	72	0	0	0	0	Lib
Eur J Biochem	72	36	12	36	12	0	12	12	Lib
FEBS Lett	52	28	28	28	28	28	0	14	Lib
Gene	105	72	24	52	40	40	16	8	Lib
J Bacteriol	108	(66)	24	24	12	0	0	0	Lib
J Biol Chem	192	180	84	72	48	12	36	12	Lib
J Gen Virol	24	(24)	(24)	12	0	0	0	0	Lib
J Mol Biol	132	60	96	36	72	24	12	12	Lib
J Mol Evol	16	(11)	6	5	0	0	0	0	Lib
J Virol	120	108	72	60	24	0	60	0	Lib
Mol Cell Biol	192	60	24	0	12	0	0	0	Lib
Mol Gen Genet	66	132	20	16	8	6	9	0	Lib
Nature	117	117	65	91	65	78	52	26	Lib
NAR	408	372	276	276	228	108	48	36	Lib
PNASU	348	348	288	144	132	144	24	12	Lib
Science	91	65	39	26	13	0	0	0	Lib
Virology	32	32	0	48	0	16	0	0	Lib
Totals	2639	2236	1415	1265	881	642	342	206	
Adjusted Totals (x85%)	2243	1901	1203	1072	749	546	291	175	

The rate at which sequence data are being published is still increasing significantly, but no longer growing exponentially. The March 1986 GenBank release contains the following numbers of papers by year of publication.

Year	1985	1984	1983	1982	1981	1980	1979	1978
In GenBank (Mar 86)	436	1195	1121	842	627	487	259	166
Estimated Papers Published	2243	1901	1203	1072	749	546	291	175
% in GenBank(Mar 86)	19%	63%	93%	79%	84%	89%	89%	95%

Similar data for the August 1986 release follow.

Year	1986	1985	1984	1983	1982	1981	1980	1979	1978
In GenBank	108	1208	1420	1175	860	660	492	268	169
Estimated Papers Published	~2700	2243	1901	1203	1072	749	546	291	175
% in GenBank	4%	54%	75%	98%	80%	88%	90%	92%	96%

Summary 1978-1985

Estimated papers published	8180
Number in August 1986 release	6694
% in August 1986 release	82%

Obviously, the flow of data has outstripped GenBank's capacity, as it has been deployed. GenBank is used in many ways. For some uses only fully annotated sequences are of value, while - at the other end of the scale - many investigators are quite ready to sacrifice annotation for timely availability of the raw sequence. This issue was discussed at the third Advisors' meeting in October 1985 in the light of the results above. To make unannotated sequences more quickly available, GenBank releases now include a new division containing the sequences being processed for the database. To increase the flow of unannotated data into the database, some of our effort has been redirected to the first stages of data entry. The 29% growth in the size of the database between the March 1986 and August 1986 GenBank releases is partly a result of this redirection of effort and the inclusion of the new unannotated division. (It is also due in great part to a growing influx of data from the EMBL Nucleotide Sequence Data Library). The unannotated entries now constitute approximately 15.5% of the entries in the database. We project that by the end of 1986 the database will be close to current, although unannotated sequences will constitute an increasing fraction of the total.

Growth of the Genbank Database (LISTED BY DR. WAYNE P. RINDONE) :

RELEASE NUMBER	ENTRIES	BASES	REPORTS	ARTICLES	MEGA- BYTES	QUARTERLY INCREASE
1. (October '82)	440	601438	591		1.06	
2. (November '82)	540	625351	696		1.42	
3. (December '82)	606	680338	778		1.58	
4. (January '83)	637	706596	819		1.71	61.8 %
5. (February '83)	705	770374	908		1.90	33.9 %
6. (March '83)	1093	1165783	1329		2.75	74.3 %
7. (April '83)	1284	1321146	1528		2.94	71.6 %
8. (May '83)	1562	1476657	1856		3.42	80.6 %
9. (June '83)	1773	1699705	2082		3.85	40.0 %
10. (July '83)	1865	1827214	2203		5.07	72.6 %
11. (August '83)	1947	1898581	2330		5.32	55.3 %
12. (September '83)	2222	2114326	2700		6.07	57.4 %
13. (October '83)	2378	2217415	2884		6.45	27.2 %
14. (November '83)	2427	2274029	2958		6.67	25.5 %
15. (December '83)	2562	2424475	3112		7.31	20.5 %
16. (January '84)	3012	2602596	3609		8.63	33.9 %
17. (February '84)	3424	2824911	4252		10.87	62.9 %
18. (March '84)	3521	2913629	4399		11.10	51.8 %
19. (April '84)	3576	2945001	4471		11.19	29.6 %
20. (May '84)	3665	3002088	4582		12.36	13.7 %
21. (June '84)	3825	3102728	4836		13.06	17.7 %
22. (July '84)	3885	3126339	4969		13.37	19.5 %
23. (August '84)	4088	3220532	5198		14.04	13.6 %
24. (September '84)	4135	3323270	5356		14.52	11.2 %
25. (October '84)	4175	3368765	5486		14.91	11.5 %
26. (November '84)	4393	3689751	5756		16.02	14.1 %
27. (December '84)	4526	3813396	6019		16.60	14.3 %
28. (January '85)	4718	4024281	6291		17.43	16.9 %
29. (February '85)	4754	4045305	6346		17.66	10.3 %
30. (March '85)	4879	4191685	6606		18.21	9.7 %
31. (April '85)	4926	4257856	6664		18.42	5.7 %
32. (May '85)	4954	4311931	6717		19.21	8.8 %
33. (June '85)	5001	4354111	6772		19.38	6.4 %
34. (July '85)	5102	4449185	6923		19.80	7.5 %
35. (August '85)	5543	5038080	7277		21.45	11.7 %
36. (September '85)	5700	5204420	7540		22.14	14.2 %
37. (October '85)	5731	5248932	7589		22.29	12.6 %
38. (November '85)	5927	5530859	7902	4697	23.39	9.1 %
39. (December '85)	6044	5667645	8074	4793	23.91	8.0 %
40. (February '86)	6646	5925429	8735	5196	25.64	9.6 %
41. (March '86)	7001	6259412	9181	5472	27.02	11.3 %
42. (May '86)	7416	6765476	9697	5768	28.62	11.6 %
43. (June '86)	8004	7416117	10355	6109	31.13	15.2 %
44. (August '86)	8832	8451024	11413	6700	33.53	22.0 %
45. (September '86)	9411	8931101	12038	7044	35.51	14.1 %
46. (November '86)	9975	9594560	12643	7342	38.64	15.2 %
47. (December '86)	10485	10301619	(estimate)		41.06	15.6 %

Genetic Sequence Data Bank
10 November 1986

GenBank(R) Release 46.0

Database Summary

9975 loci, 9594560 bases, from 12643 reports

GenBankSummary: Genetic Sequence Data Bank Summary
GenBank(R) Release 46.0 10 November 1986

Major Group of Entries	Reports	Entries	Bases
1. PRIMATE Sequences	1671	1189	1430865
2. RODENT Sequences	1775	1406	1265087
3. Other MAMMALian Sequences	322	272	276306
4. Other VERTebrate Sequences	577	491	415635
5. INVERTebrate Sequences	751	656	496184
6. PLANT Sequences	779	650	756363
7. ORGANELLE Sequences	467	400	518710
8. BACTerial Sequences	1224	834	1140574
9. Structural RNA Sequences	750	653	70752
10. VIRAL Sequences	1823	1138	1626279
11. BacterioPHAGE Sequences	346	167	279907
12. SYNTHETIC Sequences	275	239	74018
13. UNANNOTATED Sequences	1883	1880	1243880
Summary:	12643	9975	9594560

Sequences With Greater Than 30,000 bp

LOCUS	GROUP	BP
HUMTPA	PRIMATE	36594
HUMFIXG	PRIMATE	38059
HUMHBB	PRIMATE	73360
ADBCG	VIRAL	35937
PT7	PHAGE	39936
LAM	PHAGE	48502
EBV	EBV	172282

Primate Summary: Genetic Sequence Data Bank Summary
GenBank(R) Release 46.0 10 November 1986

Code		Source of Sequence	Reports	Entries	Bases
1.	AGM	African Green Monkey	15	14	8327
2.	ATR	Owl Monkey	3	3	791
3.	BAB	Baboon	5	5	3944
4.	CEB	Cebus Monkey	2	2	190
5.	CHP	Chimpanzee	10	10	14839
6.	COL	Colobus polykomos Monkey	2	2	1494
7.	GCR	Galago crassicaudatus (bush baby)	28	28	6044
8.	GIB	Gibbon	3	2	1791
9.	GOR	Gorilla	1	1	2151
10.	HUM	Human	1585	1105	1384497
11.	LEM	Lemur	1	1	1786
12.	MAC	Rhesus Monkey	5	5	1664
13.	MNK	Monkey	8	8	1182
14.	ORA	Orangutan	3	3	2165
		Summary:	1671	1189	1430865

Genetic Sequence Data Bank
3 September 1985

GenBank(R) Release 36.0

Database Summary

5700 loci, 5204420 bases, from 7540 reports

GENBANKSUMMARY: GENETIC SEQUENCE DATA BANK SUMMARY
GENBANK(R) RELEASE 36.0 3 September 1985

Major Group of Entries	Reports	Entries	Bases
1. PRIMATE Sequences	992	692	780200
2. RODENT Sequences	1188	948	729086
3. Other MAMMALian Sequences	233	198	169397
4. Other VERTebrate Sequences	416	362	253855
5. INVERTebrate Sequences	544	479	306504
6. PLANT Sequences	465	405	396878
7. ORGANELLE Sequences	375	315	387867
8. BACTerial Sequences	817	565	687371
9. Structural RNA Sequences	611	524	57751
10. VIRAL Sequences	1415	892	1171680
11. BacterioPHAGE Sequences	277	138	207668
12. SYNTHETIC Sequences	207	182	56163
Summary:	7540	5700	5204420

Primate SUMMARY: GENETIC SEQUENCE DATA BANK SUMMARY
GENBANK(R) RELEASE 36.0 3 September 1985

Code	Source of Sequence	Reports	Entries	Bases
1. APE	Ape	6	6	5529
2. GCR	Galago crassicaudatus (bush baby)	19	19	4183
3. HUM	Human	932	632	752495
4. LEM	Lemur	1	1	1786
5. MNK	Monkey	34	34	16207
	Summary:	992	692	780200

APPENDIX

PART C

"THE HUMAN GENE MAP"

BY R.S. SPARKES

3

The human gene map

Robert S. Sparkes

Introduction
History
Methods of gene mapping
  Genetic linkage; association
  Somatic cell hybridization
  Gene dosage
  In situ hybidization
  Chromosome sorting
  Comparative mapping
Human chromosome fragile sites
Gene mapping in human cancer
Clinical human gene map
Useful clinical mapping relationships
Future developments
Glossary
References

INTRODUCTION

The assignment of human genes to specific chromosomes and parts of chromosomes is one of the most rapidly advancing areas in human genetics. Over the past several years there has been a number of reviews covering human gene mapping. Renwick (1971) summarized its history and status up to that time. In 1977 McKusick and Ruddle reviewed the status of the map and the methodologies used for mapping. McKusick (1980a, 1980b, 1982) has subsequently published several follow-up articles, which have a clinical orientation. Siniscalco (1979) made an extensive review of the approaches to human gene mapping. Conneally and Rivas (1980) focused on the linkage aspects of mapping in their review. Ruddle (1981) summarized the application of recombinant DNA and somatic cell methodologies to gene mapping. Shows, Sakaguchi and Naylor (1980) reviewed human gene mapping with a focus on the recombinant DNA approaches. Kao (1983) discussed the use of somatic cell genetics in gene mapping. In addition, International Workshops on Human Gene Mapping are held every two years, the latest being convened in Los Angeles and the proceedings have recently been published (Human Gene Mapping VII, 1984). These meetings critically review old and new gene mapping information and summarize accepted gene maps and gene nomenclature. This serves as the main reference for human gene mapping until superseded by the proceedings from the next workshop. References to the prior workshops can be found in the latest published proceedings (Human Gene Mapping VII, 1984).

HISTORY

Because of the characteristic pattern of X-linked inheritance, the first assignment of a gene for color blindness to the human X chromosome was possible in 1911 (Wilson, 1911). But it was not until 1951 that the first autosomal linkage was demonstrated for Lutheran blood group and secretor status (Mohr, 1951). In 1968 the first assignments to specific human autosomes were made, and by 1976 at least one gene had been assigned to each autosome. The rapid recent progress has been due in large measure to the convergence of a number of developments including: improved statistical analyses combined with computer capabilities; long-term storage of cells, cell cultures and body fluids in liquid nitrogen; identification of each chromosome by specific banding techniques; development of new phenotypic and DNA polymorphisms; and somatic cell hybridization techniques.

METHODS OF GENE MAPPING

Tables 3.1 and *3.2* list the significance of gene mapping studies in relation to clinical and basic considerations respectively. As can be seen, there are a number of human biological problems which can benefit from gene mapping information; and they range from practical clinical applications, such as in genetic counseling, to more

Table 3.1 Clinical significance of gene map information

(1) Identification of genetic heterogeneity (genocopies) – implications for diagnosis and genetic counselling
(2) Identification of persons at risk for a genetic disease in the presymptomatic stage by linkage study
(3) Linkage for prenatal diagnosis of trait which is not expressed at early stage of development or in tissue available for testing
(4) Carrier detection in recessive conditions in which direct expression of affected gene cannot be detected
(5) Identify parent in whom a chromosome abnormality originated in a child – implications for eventual recognition of factors causing chromosome abnormalities with future prevention of other affected children
(6) Insight into pathogenesis of abnormal phenotype with chromosome abnormalities – ? imbalance of specific genes on affected chromosome

Table 3.2 Basic biological significance of gene map information

(1) Increase knowledge of factors related to genetic recombination – implications for evolution
(2) Identify factors which may affect crossing-over (e.g. sex, age, race, heterochromatin, centromere effects)
(3) Do physical map distances on a chromosome correlate directly with crossing-over frequencies?
(4) How expression of a gene is affected by its location (position effect)
(5) Improve understanding of chromosome structure
(6) Relation of chromosome bands to gene content and function
(7) Are gene locations related to their function – e.g. expression of genes in families or clusters, or relation to metabolic functions
(8) Identification of genetic control and regulatory functions
(9) Insight into relation of neoplasia and associated chromosome changes
(10) Comparison of gene maps between species and relation to evolution
(11) Genetic interaction
(12) Redundancy of genetic information
(13) Presence and role of pseudogenes
(14) Effect of heterochromatin on gene expression

basic disciplines including human evolution. About 15 years ago a number of factors converged which made it technically feasible to carry out human gene mapping analyses (*see above*). As a result, a number of methodologies have been developed for these studies. *Table 3.3* lists those methods which have been most successful in mapping. Some of these will now be discussed in more detail.

Table 3.3 Methods for gene mapping

(1) Linkage using polymorphic DNA (RFLPs) and phenotypic and chromosome markers
(2) Interspecific somatic cell hybridization with DNA and phenotypic markers
    (a) microcell mediated gene transfer
    (b) chromosome mediated gene transfer
    (c) radiation-induced gene segregation
    (d) DNA mediated gene transfer
(3) Gene dosage
(4) *In situ* DNA hybridization to banded chromosomes
(5) Chromosome sorting
(6) Amino acid sequence of protein resulting from gene deletion and fusion
(7) Linkage homology between species – comparative mapping
(8) Inheritance pattern for X-linkage
(9) Centromere mapping (teratomas)

Genetic linkage; association

The study of genetic linkage in families was the earliest successful approach for gene mapping in humans and has experienced a renaissance with the development of restriction fragment length polymorphisms (RFLPs) of DNA (*see below*). According to Mendel's law of independent segregation or assortment, transmission of genes from parents to children occurs at random. However, when genes are linked, this law does not hold. Linkage is the tendency of genes on the same chromosome to segregate together, i.e. linked genes are transmitted to the same gamete more than 50% of the time. The closer genes are to each other, the more frequently they will be transmitted together. This is related to crossing-over that occurs in meiosis and results in the reshuffling of genes between homologous chromosomes. Thus, genes that are far apart on the larger chromosomes may segregate randomly in relation to each other. All genes that are located on the same chromosome are said to be syntenic (Renwick, 1971); all genes that are linked are syntenic, but the converse may not be true.

Because of the distinct inheritance pattern for X-linked genes it is possible to assign genes to the X chromosome through the X-linkage pattern. Assignment of a gene to a specific autosome by linkage is not so easy because there are 22 pairs of autosomes. Thus, establishment of an autosomal linkage will not assign a gene to a specific autosome, unless one of the linked genes has been previously assigned to that autosome.

Attempts are being made to relate crossing-over to the physical lengths of DNA and chromosomes. The unit of distance, a Morgan, is defined as the length of chromosomal segment which, on average, experiences one cross-over per strand of DNA. In actual linkage studies one generally deals in centimorgans (cM), which is 1/100 Morgan and also represents 1% recombination (generally represented, θ =

0.01). Crossing-over is more common in the homogametic sex, which in humans is the XX female; there is no clear explanation for this sex difference. Because of this difference between the sexes, Renwick (1971) has suggested a neuterized map length of 33 morgans. Ruddle (1981) estimates that in an 850 band human karyotype each band contains, on average, three centimorgans; and based on 3000 centimorgans in humans, each centimorgan contains 10^6 nucleotide base pairs.

To carry out genetic linkage studies in humans one has to utilize genetic variation. Common genetic variations are called polymorphisms and the polymorphic markers which have been most widely used in humans are red blood cell antigens, white blood cell antigens (HLA), isoenzymes and serum proteins. Chromosomal variants can also be used. But the area which has seen the most rapid growth and has the greatest potential for continued expansion is the restriction fragment length polymorphisms of DNA.

Desirable characteristics for linkage studies include: dominantly or co-dominantly expressed traits; expression of a trait in an easily accessible tissue; complete penetrance of the gene; early onset of the phenotypic expression; lack of genetic heterogeneity; and availability of highly polymorphic traits.

Linkage analysis has certain advantages over other approaches to gene mapping. These include the ability to study traits that are not manifest at the single cell level. It is also useful for measuring distances between loci and can give a linear map of gene order.

Limitations to linkage analysis in man compared to other organisms include: uncontrolled matings; family size tends to be relatively small with a limited number of generations available for study; there is a long generation time; genetic heterogeneity may be unrecognized; problems of non-paternity can lead to errors in interpretation of results; there may be lack of cooperation of important persons in a pedigree; and there may be ethical and legal problems in trying to carry out the necessary studies for linkage.

Linkage analysis may be direct when the parental genotypes for two loci are known, both for which alleles are present and in what phase they are in relation to each other (whether they are on the same or homologous chromosomes). In this instance, linkage can be evaluated by counting the number of recombinant and non-recombinant types in the children of such parents. Unfortunately, this is not the usual situation for linkage analysis in humans. As a result, more complicated statistical approaches have been developed, as described below.

These approaches are called indirect analyses because the genotypes of certain persons, such as parents and sibs, may not be fully known in relation to allelic types and their phase. To handle these problems, statistical approaches have been developed which are especially useful for two-generation families (Conneally and Rivers, 1980). The first attempt at a statistical approach to linkage used the y statistics of Bernstein. Subsequently, Penrose developed the sib–pair method. More recently the method of likelihood ratio or odds of Haldane and Smith was devised. This method determines the amount of information available in a collection of data on two loci and compares the probability of obtaining such data if the two loci are linked with the probability if they are not linked. This ratio of the two probabilities gives the odds for or against linkage. Morton (1955) developed the sequential analysis method in which the use of logarithms also makes it easier to combine results from different families. The term *lod* is for 'log odds' which stands for the log probability ratio. The log 10 is generally used to determine the lod score. By convention, a lod score of +3 or more indicates linkage, while a lod score of −2 or less rules out linkage. The lod score can be calculated for different recombination values of θ between 0.0 and 0.5. The recombination giving the highest lod score is used to give the maximum likelihood of linkage. It is of interest to consider why the maximum recombination fraction is 0.5. The basis for this is that for each crossing-over only 50% of the offspring will be recombinants and 50% will be of the original non-crossover type. This results from the fact that crossing-over occurs during the four strand stage of meiosis and that only two of the strands are involved in the crossing-over. Thus, the maximum recombination is 0.5.

Ott (1974, 1977) developed a computer program called LIPED to facilitate the handling of the data and the statistical calculations. This is the method most widely used at the present time, but attempts are continuing to improve the effectiveness of the analysis as well as to make it more efficient.

Because of the recent development of the RFLPs of DNA, and the likelihood that they will become the mainstay of linkage analysis in the future, a few additional comments will be made regarding their use. All genetic variation reflects differences at the DNA level. This is true for the phenotypic markers which reflect genetic changes based upon differences between alleles that can be determined by electrophoretic techniques, antigenic differences or other biochemical parameters. With the RFLPs of DNA, the differences are detected directly in the DNA. The study of these DNA polymorphisms is based upon the following. Restriction endonucleases split DNA at specific base sequences. Because these base sequences may vary between persons, different size DNAs are produced by the same enzyme. By the use of different restriction endonucleases, a potentially large number of different base sequences may be detected in a population. By isolating DNA from cells of individuals and treating them with restriction endonucleases, the DNA fragments can be separated by molecular size using agarose gel electrophoresis. One then hybridizes radioactive-labeled DNA probes of specific genes or DNA probes of non-specific or unidentified functions to the DNA which has been treated with the restriction endonucleases and separated by the electrophoresis. Differences in size of the DNA segments can be detected with the probes. These differences can then be treated in the standard linkage approach described above. With the availability of a large number of restriction endonucleases, it is very likely that each and every gene will be found to demonstrate polymorphisms following treatment with one or more of these restriction endonucleases. It has been estimated that perhaps only 165 DNA probes could span or map the human genome (Botstein *et al.*, 1980). This would be true if the probes were mapped at random. However, it has been estimated that many more markers may have to be studied and evaluated in order to find the minimum number which would be randomly distributed over the genome (Lange and Boehnke, 1982). Nevertheless, it appears at this time, that with sufficient effort, the appropriate probes can be isolated and used to map eventually the whole human genome.

Genetic linkage and association are terms often used interchangeably, but they describe different phenomena with different implications (Vogel and Motulsky, 1979; Hodge and Spence, 1981). Within a single family it may not be possible to distinguish linkage from association. Evaluation of each requires the use of genetic polymorphisms, which may in part contribute to the confusion. Linkage relates to genetic loci, and the alleles at these loci are useful only as markers for the loci. The evaluation of linkage requires the study of families, especially large families.

Association refers to a concurrence greater than predicted by chance between a specific allele and another trait which may or may not have an obvious genetic basis. Evaluation of association requires the study of unrelated individuals. Association may given an insight into susceptibility and pathogenesis of a trait or disease. Thus, association studies may be used to identify a genetic factor in a disease. With the exception of linkage disequilibrium (*see below*), association is not due to linkage. Association may result from: pleiotrophic effects of a single gene; epistatic interaction; selection in relation to environmental factors (e.g. sickle cell gene and malarial resistance); and population stratification.

Many early examples of association were related to the ABO blood groups, because this was the earliest polymorphic trait available for study. When the highly polymorphic HLA system became available, it was applied to a large number of diseases in search for associations. Probably the best known is the association of HLA B27 antigen with ankylosing spondylitis.

In addition to the association of the sickle cell gene and resistance to malaria, a similar resistance has been noted for glucose 6-phosphate dehydrogenase (G6PD) deficiency and malaria. Another association that has been described in relation to malaria is with the Duffy blood group system in which individuals of the Fy (a−b−) phenotype demonstrate resistance to malaria. Evidence has been presented that the Duffy a and b antigens may act as red cell receptors for the malarial parasites. As expected, the genotypes for the resistance to malaria have been found to occur with a high frequency in malarial areas. These are also nice examples of how environmental factors may contribute to specific allelic frequencies.

Demonstration of associations may help to: identify a genetic component to a disease; clarify pathogenesis of a disease; facilitate preclinical testing to identify persons at risk so that appropriate counseling may be given; and recognize disease heterogeneity.

Above, it was noted that linkage disequilibrium may contribute to an apparent association; this has sometimes been referred to as 'hitch-hiking'. This is said to occur when specific alleles at two different loci are found together more than expected. This may result from natural selection, or by chance, such as when two populations have recently been mixed without time for randomization of the genotypes in relation to each other. Examples are seen where it is clear that an association is present between two different genetic loci; with the HLA system, this has been suggested for insulin-dependent diabetes mellitus, hemochromatosis and congenital adrenal hyperplasia due to 21-hydroxylase deficiency.

Somatic cell hybridization

The second method and the one which has proved to be the most effective and successful in mapping human genes is the use of interspecific somatic cell hybrids. This is a parasexual method that can bypass sexual reproduction. The basic method for human gene mapping is to fuse human somatic cells (e.g. lymphocytes, fibroblasts) with an established rodent cell line, such as from mouse and Chinese hamster. Most often, the rodent cell line has a deficiency in an enzyme, such as HPRT, which makes it possible to select against the rodent parent cell line in culture and also to permit selection for the interspecific somatic hybrid cells. If human lymphocytes are used these can be easily physically removed from the tissue culture following fusion because they will not attach to the tissue culture flask. If human fibroblasts are used, the parental cells can be selected against through the use of ouabain. The fused hybrid cells can then be grown in tissue culture. This results in the loss of human chromosomes from the somatic cell hybrids. The loss of the human chromosomes is approximately random so that cloning of the somatic cell hybrid cells gives hybrid cell clones with different human chromosome contents. By correlating the expression of human genes or the direct measurement of the human genes (DNA), the genes can be mapped to specific human chromosomes. This method has been particularly powerful because it does not depend upon the need for polymorphic human markers, but instead uses species differences to distinguish the human genes from the rodent genes. For many of the phenotypic markers, enzyme electrophoresis has been most widely used.

This approach has been greatly expanded recently through the use of the DNA probes. One advantage of the direct DNA studies is that it is independent of the need for gene expression. Indeed, one can map DNA probes even though their specific function is not known. The somatic cell hybrid strategy also can map genes to a given chromosome, i.e. to establish synteny. By using human cells with chromosome rearrangements, such as translocations, genes can be regionally mapped to chromosomes, again by correlating the presence of or expression of a human gene with the part of the human chromosome which remains in the somatic cell hybrid. This general approach has been supplemented by refinements and modifications. For example, microcell mediated gene transfer can facilitate the formation of somatic cell hybrids containing a single human chromosome or very few human chromosomes. With this technique one fuses microcells which contain one or a few human chromosomes with the rodent cell line. Second, with the isolation of specific chromosomes, as through the use of chromosome sorting, it is possible to have cells take up isolated chromosomes by endocytosis or by direct microinjection of human chromosomes into a rodent cell. Another approach which has been utilized for regional mapping is to irradiate the human cells before fusion with the rodent cells. The radiation breaks the human chromosomes and, following fusion, only parts of the chromosomes will be retained. Particularly through the use of a selectable marker, it is possible to see which human genes are present together in the somatic cell hybrid following infusion.

Gene dosage

A third general approach for human gene mapping is based upon the gene dosage concept. This method has been used primarily for mapping genes for enzymes, and it is also possible to measure the amount of allelic DNA present in a cell. It is based upon the concept that the amount of expressed gene product directly reflects the number of alleles which contribute to that gene product, such as enzyme activity. This holds true for a number of enzymes, particularly those found in red blood cells. Thus, it was found that the superoxide dismutase-1 activity in trisomy 21 is approximately one and one-half times normal, indicating that the genetic locus for this enzyme is on chromosome 21 (Sinet *et al.*, 1974). In a similar fashion, it has been possible to map regionally the locus for esterase D by studying patients with partial deletion of the long arm of chromosome 13 (Sparkes *et al.*, 1980).

In situ hybridization

A relatively new technique may prove to be very powerful for mapping, and particularly regional mapping, of genes. This is the *in situ* hybridization method in which a labeled DNA probe is hybridized directly to chromosome spreads and a search is made for the location of the labeled probe on the human chromosomes. This has proved to be very effective for multi-copy genes but has some limitations for single copy genes; the technique may need to be improved before its full value can be realized for mapping single copy genes.

Chromosome sorting

Another approach to gene mapping is to use separated chromosomes. Somatic cell hybridization, as described above, segregates human chromosomes, but on a rodent background. A fluorescence activated cell sorter (FACS) can physically separate chromosomes from dividing human cells (Lebo, 1982). First, a suspension of metaphase chromosomes is made. This chromosome suspension is then passed through the FACS. The chromosome separation depends upon the use of DNA stains and the power of the laser to excite the stain. For example, ethidium bromide stains all DNA base pairs non-specifically, while Hoechst 33258 preferentially stains AT base pairs and chromomycin A3 stains GC base pairs. Powerful lasers are needed to separate chromosomes with these latter stains. Chromosomes may be resolved into as many as 21 peaks through the combination of these two stains and through the use of two consecutive lasers. It is anticipated that with continued technical improvements, separations will be improved as well.

The separated chromosomes can be used for gene mapping as follows. This approach uses a radioactively labeled DNA probe which is to be mapped. The DNA in the separated chromosomes can be enzymatically digested and the probe hybridized to this. If hybridization occurs, the gene is said to be mapped to the chromosomes from which the DNA was extracted. In theory, if each chromosome could be separated from all the others, the mapping would have been accomplished. However, in practice it seems that it is more practical to separate chromosome fractions containing more than one chromosome type; a positive hybridization indicates that the DNA probe is mapped to one of the chromosomes in the fraction. Next, chromosome translocations containing those chromosomes in the chromosome fraction are treated in the same fashion. The mapping is accomplished by identifying hybridization with a chromosome fraction that did not previously contain the rearranged chromosome segment.

This strategy can also be used to make chromosome specific DNA libraries, which then can be used for molecular mapping.

The future use of the chromosome sorting strategy will depend upon improved separations and the wider availability of facilities to separate the chromosomes. If improvements can be made in these areas, this approach may achieve a more prominent role in mapping of isolated genes or DNA.

Comparative mapping

Comparative mapping between species can also give useful insights into mapping of human genes. Particularly among mammals, there has been evolutionary conservation of many syntenic genes. Thus, if one finds that two genes are located on the same chromosome in a rodent, such as the mouse, it is very likely that the same genes will be linked or syntenic in humans. There are at least 23 conserved linkage groups which have already been demonstrated between mouse and man. The reasons for this evolutionary conservation are not clear but raise questions as to possible functional relationships.

HUMAN CHROMOSOME FRAGILE SITES

Human chromosome fragile sites are defined as specific points or loci which have the following characteristics (de la Chapelle and Berger, 1984): tendency to form chromosome and chromatid gaps; are observed at the same locus in any patient or kindred; are inherited in a Mendelian co-dominant fashion; and result in abnormal chromosomes such as deletions, multiradial figures, acentric fragments, by breakage at the fragile site under appropriate *in vitro* conditions. Because of these features, the fragile sites have the potential to be used for genetic linkage analysis. The currently recognized fragile sites in human chromosomes are summarized in *Table 3.4*. They can be characterized by the conditions which induce the fragile sites. The folate sensitive fragile sites are expressed in tissue culture in which there is a relative deficiency of thymidylate, which can be accomplished through a number of different procedures. The second group requires the presence of distamycin A in the tissue culture medium to produce these sites. There is only one fragile site which requires the presence of bromodeoxyuridine (BrdU) in the culture medium for the site to be produced. Finally, there is a small group of common fragile sites which are usually expressed in only a small proportion of the metaphases, but they appear to be present in a high frequency and possibly in all humans.

Table 3.4 Human chromosome fragile sites

Type	*Location*
Folate sensitive	2q11.2, 2q13, 6p23, 7p11.2, 8q22.3, 9p21.1, 9q32, 10q23.3, 11q13.3, 11q23.3, 12q13.1, 16p12.3, 20p11.23, Xq27.3
Distamycin A inducible	16q22.1, 17p12
Bromodeoxyuridine requiring	10q24.2
Common sites	3p14, 6q26, 16q23

GENE MAPPING IN HUMAN CANCER

It has been recognized for a number of years that chromosome changes are commonly, and perhaps always, found in cancer cells. Particularly through the use of chromosome banding techniques, it is clear that these chromosome changes are non-random (Mitelman, 1983; de la Chapelle and Berger, 1984). Although the basis of these chromosome rearrangements is poorly understood, there is a growing awareness that the chromosome changes may be related to specific genes. For example, in the Philadelphia chromosome of chronic myelogenous leukemia there is evidence that the abl oncogene, which is usually found on the long arm of chromosome 9 (9q34), is translocated to the region of the lambda light chain immunoglobulin genes on chromosome 22. With Burkitt lymphoma, chromosome translocations involving band 8q24 (site of myc oncogene) and 2p11-13 (site of kappa light chain immunoglobulin gene) or 14q32 (site of heavy chain immunoglobulin gene) or 22q11 (site of lambda light chain immunoglobulin gene) have been described. Other examples are rapidly accumulating.

Table 3.5 Human oncogene maps

Oncogene	*Chromosome*
nras	1p31.1 – p22.1 or 1p13
blm1	1p32
sk	1q12 – qter
fos	2
nmyc	2
raf1	3p25
raf2p	4
fms	5q34
myb	6q15 – q24
kras1	6p23 – q12
erbb	7pter – q22
mos	8q22
myc	8q24
abl	9q34
hras1	11p15.1 – p15.6
kras2	12p12 – q24.2
fes	15q25 – q26
erba1	17p11 – q21
src	20
sis	22q12.3 – q13.1
hras2	X

Rapid progress is being made in the mapping of the oncogenes. *Table 3.5* summarizes current mapping information on 21 of these genes. Although the relationship between oncogenes and human cancers is not fully understood, work to date has given interesting and provocative leads to the study of human cancers, such as noted above, in their relationship to chromosome rearrangements in cancers. There is growing evidence that the number of oncogenes or the amount of oncogene product is increased in some tumors. This has suggested that the cellular mechanism for the expression of oncogenes in producing a tumor appears to be dominant. The history of the gene mapping of the retinoblastoma gene is reviewed next. This illustrates the application of a number of different genetic techniques for mapping and also gives an insight into possible tumorigenesis through a recessive mechanism.

Retinoblastoma is a developmental eye tumor which has an hereditary basis in about 40% of cases. This estimate is based upon: autosomal dominant inheritance with 80–90% penetrance in some cases; all bilateral cases are hereditary; and 15–20% of unilateral cases are hereditary. The remaining 60% of the cases occur sporadically. A small group of cases has a constitutional chromosome deletion of band 13q14. Earlier studies by somatic cell hybridization had mapped the gene for esterase D enzyme to chromosome 13. Through the study of the patients with a 13q14 chromosome deletion with retinoblastoma, we were able to map regionally the locus for esterase D to band 13q14 (Sparkes *et al.*, 1980). This suggested that the retinoblastoma genetic factor which is deleted in the chromosome deletion form of retinoblastoma is closely linked to the esterase D locus.

Family studies demonstrated close linkage between the gene for the inherited form of retinoblastoma and esterase D (Sparkes *et al.*, 1983). This close linkage suggested that a chromosome deletion could be inferred from esterase D activity. The study of one such patient with retinoblastoma and apparent normal chromosomes 13 demonstrated only half normal esterase D activity, suggesting a chromosome deletion too small to be seen by standard techniques (Benedict *et al.*, 1983). Chromosome studies on the tumor from this patient demonstrated only one chromosome 13 present and there was no esterase D activity. These results suggested that the remaining chromosome 13 was the one with the presumed submicroscopic deletion. From the results in this patient we postulated that the tumor expression is related to a recessive mechanism in the retinoblastoma cells.

Godbout *et al.* (1983) extended the esterase D studies in patients with retinoblastoma. They found that in some patients who are heterozygous for the electrophoretic variants of esterase D in their normal tissues, the tumors expressed only one allele form of the esterase D. They suggested that there may be inactivation of one of the esterase D alleles. Recombinant DNA studies by Cavenee *et al.* (1983) afforded alternative explanations. These investigators isolated DNA probes from chromosome 13. Then, through the use of restriction endonucleases, they found restriction fragment length DNA polymorphisms for these probes. They identified for which of the probes the patient's normal cells were heterozygous. They then evaluated the tumor cells for the presence of these DNA segments and found that some of the tumors had only a homozygous make-up for these fragments. They interpreted their results to indicate that the dominant normal allele at the retinoblastoma locus had been lost in the formation of the tumor. In one case this could be attributed to loss of one chromosome 13. Other possibilities leading to the presence of a homozygous state for the mutant retinoblastoma allele included: non-disjunction with loss of the normal chromosome 13 and reduplication of the chromosome containing the retinoblastoma allele; mitotic recombination; gene conversion; gene or small chromosome deletion of the normal allele; and a point mutation of the normal allele. This study clearly emphasizes the utility of the DNA techniques for better interpretation of possible mechanisms leading to chromosome imbalance. As these authors suggested, this method could potentially be used to map cancer genes to chromosomes as follows. If one has one or more gene markers (DNA restriction fragment length polymorphisms or phenotypic markers) for each chromosome, one could potentially determine for which chromosome a tumor becomes homozygous. The responsible gene for the cancer could then be assigned to that chromosome.

CLINICAL HUMAN GENE MAP

More than 800 human genes and/or DNA probes have been mapped. This is too extensive a list to include here; it is summarized in Human Gene Mapping VII (1984). I have limited the presentation of assigned genes to those focusing on clinical genetics and human disease (disadaptive condition).

Significant advances have been made in the mapping of the genes for human diseases. *Table 3.6* summarizes most of the disease genes which have been mapped. It includes some genes that have been mapped only provisionally and these need confirmation before they can be accepted; but it is very likely that many or most of these will be confirmed by subsequent studies. For many of the diseases listed, it is thought that the mutations have affected structural genes, but in most instances direct demonstration of this by appropriate DNA studies has not yet been accomplished. While the list in *Table 3.6* is impressive, it should be remembered that this still represents only a very small proportion of the total of known single gene human diseases which are listed by McKusick (1983), and in turn this catalog represents only a small proportion of the potentially larger number of hereditary diseases based upon a postulated 50000 genes in man. Many of the disease genes which have been mapped have used somatic cell hybridization to map an enzyme which is known to be affected in a genetic disease. A few dominant disorders have been mapped by linkage studies. As with mapping in general, the future advances in mapping of human hereditary diseases will come through the application of the recombinant DNA techniques.

USEFUL CLINICAL MAPPING RELATIONSHIPS

Gene mapping information already has several clinical applications and implications. Presymptomatic diagnosis of late onset dominant disorders may be made through genetic linkage. Persons at risk of developing an autosomal dominant disorder may be identified through linkage and, if available, preventive measures may be instituted to delay or prevent the full clinical expression of the disorder. Genetic linkage may also be helpful in prenatal diagnosis, as for myotonic distrophy (Schrott, Karp and Omenn, 1973) and congenital adrenal hyperplasia due to 21-hydroxylase deficiency (Pollack *et al.*, 1979). Other specific diseases in which this is currently possible are summarized in *Table 3.7*. It is anticipated that this table will rapidly expand with the application of the DNA polymorphisms with linkage analysis.

Recent advances in the linkage analysis of the dominant disorder, Huntington's disease, using DNA probes has made presymptomatic diagnosis very likely in some if not all affected families (Gusella *et al.*, 1983).

This gene mapping information also has the potential for detection of recessive problems in which the expressed gene product is either unknown or undetectable. An example of this is phenylketonuria (Woo *et al.*, 1983). The detection of DNA RFLPs for the enzyme phenylalanine hydroxylase (which is deficient in phenylketonuria), has made it possible to identify heterozygous carriers in some families. The same DNA probe can be used for prenatal diagnosis of phenylketonuria in informative families. This information can be applied to genetic counseling and improved management of the genetic problem in this and other recessive problems.

Probably the most extensive experience in prenatal diagnosis using DNA probes has been with inherited hemoglobin problems. Molecular genetic linkage studies of these diseases were first applied in 1978 (Kan and Dozy, 1978). An RFLP was found to be frequently associated with the sickle cell gene in black patients. In informative families, prenatal diagnosis was possible through the evaluation of the DNA from amniotic fluid cells. This obviated the need for fetal blood sampling which carries a relatively high risk to the pregnancy.

Table 3.6 Genetic diseases which have been mapped†

Disease	*McKusick** number*	*Affected gene product*	*Locus symbol*	*Chromosome assignment*
Amino acid metabolism				
Argininosuccinic aciduria	20790	Argininosuccinate lyase	ASL	7p21–q22
Atypical phenylketonuria*	26163	Quinoid dihydropteridine reductase	QDPR	4
Citrullinemia	21570	Argininosuccinate synthetase	ASS	9q34–qter
Homocystinuria*	23620	Cystathionine β-synthase	CBS	21
Ornithine transcarbamoylase deficiency	31125	Ornithine carbamoyltransferase	OTC	Xp21
Orotic aciduria-1*	25892	Orotate phorphoribosyltransferase and OMP decarboxylase	URIC	3 cen–q21
Phenylketonuria*	26160	Phenylalanine hydroxylase	PAH	12
Tetrahydropteroyl-glutamate methyltransferase deficiency	15657	5-Methyltetrahydropterol-L-glutamate: L-homocysteine s-methyltransferase	MTR	1
Carbohydrate metabolism				
Fucosidosis	23000	α-Fucosidase	FUCA1	1p34
Galactokinase deficiency	23020	Galactokinase	GALK	17q21–q22
Galactose-4-epimerase deficiency	23035	Galactose-4-epimerase	GALE	1pter-p32
Galactosemia	23040	Galactose-1-phosphate uridylyltransferase	GALT	9p21–p13
Mannosidosis	15458	α-Mannosidase-B	MANB	19pter–q13
Pompe's disease	23230	α-Glucosidase, acid	GAA	17q22–q25
Lipid metabolism				
Fabry's disease	30150	α-Galactosidase	GLA	Xq21–q24
Gaucher disease	23080	Acid β-glucosidase	GBA	1q
GM_1 gangliosidosis	23050	β-Galactosidase 1	GLB1	3p21–cen
GM_2 gangliosidosis (Tay-Sachs)	27280	β-Hexosaminidase-A	HEXA	15q22–q25.1
GM_2 gangliosidosis (Sandhoff's)	26880	β-Hexosaminidase-A and -B	HEXB	5q13
Metachromatic leucodystrophy	25010	Arylsulfatase-A	ARSA	22q13.31–qter
Wolman's disease	27800	Acid lipase	LIPA	10
Mucopolysaccharide metabolism				
β-Glucuronidase deficiency (MPS VII)	25322	β-Glucuronidase	GUSB	7cen–q22
Hunter's disease	30990	Sulfoiduronate sulfatase	SIDS	X
Maroteaux-Lamy syndrome	25320	Arylsulfatase-B	ARSB	5
Mucopolysaccaridosis I*	25280	Alpha-L-iduronidase	IDUA	22pter–q11
Lipoprotein metabolism				
Combined apo AI-CIII deficiency*	10768	Apo AI	APOA1	11p–q13
	23455	Apo CIII	APO C3	11p–q13
Dysbetalipoproteinemia	20776	Apolipoprotein E	APOE	19
Familial hypercholesterolemia	14389	LDL receptor	FHC	19
Norum disease	24590	Lecithin-cholesterol acyltransferase	LCAT	16q21–q22
Nucleic acid metabolism				
Hyperuricemia (urolithiasis)	10260	Adenine phosphoribosyl transferase	APRT	16q12–q22
Lesch-Nyhan syndrome	30800	Hypoxanthine phosphoribosyl transferase	HPRT	Xq26–q27
Other enzyme deficiencies				
Acatalasia	11550	Catalase	CAT	11p13
Acid phosphatase deficiency (lysosome storage disease)	17165	Acid phosphatase-2	ACP2	11p12–cen
Hemolytic anemia				
Adenylate kinase deficiency	20160	Adenylate kinase	AKI	9q34
Glutathione peroxidase deficiency	23170	Glutathione peroxidase-1	GPX1	3p13–q12
Glucosephosphate isomerase deficiency	17240	Glucosephosphate isomerase	GPI	19p13.2–q13
G6PD deficiency	30590	Glucose-6-phosphate dehydrogenase	G6PD	Xq28
Glutathione reductase deficiency	23180	Glutathione reductase	GSR	8p211
Hexokinase deficiency	23570	Hexokinase-1	HK1	10
Phosphogluconate dehydrogenase deficiency	17220	6-Phosphogluconate dehydrogenase	PGD	1pter–p36.13
Phosphoglycerate kinase deficiency	31180	Phosphoglycerate kinase	PGK	Xq13
Triosephosphate isomerase deficiency	27580	Triosephosphate isomerase-1	TPI1	12p13
Heme and hemoglobin				
Acute intermittent porphyria	17600	Uroporphyrinogen 1 synthase	UPS	11q23–qter
Hemoglobinopathies (sickle cell and thalassemias)	14180	α-Chain	HBA	16pter–p12
	14190	β-Chain	HBB	11p15
Methemoglobinemia	25080	NADH-diaphorase	DIA1	22q13.31–qter
Blood group				
Rh incompatibility	11170	Rhesus blood group	Rh	1p36–p34

Table 3.6 (continued).

Disease	*McKusick*** *number*	*Affected gene product*	*Locus symbol*	*Chromosome assignment*
Other				
Elliptocytosis	13050		ELI	1pter–p34
Serum abnormalities				
Analbuminemia	10360	Albumin	ALB	4q11–q22
Antithrombin III deficiency	10730	Antithrombin III	AT3	1q23
Clotting factor VII deficiency*	22750	Factor VII (proconvertin)	F7	13q34
Clotting factor XIII deficiency*	13457	Factor XIII, A component	F13A	6
Clotting factor X deficiency*	22760	Factor X (Stuart-Power factor)	F10	13q34
Clotting factor XII (Hageman) deficiency*	23400	Factor XII	F12	6pter–p23
Dysfibrinogenemia	12482,3,5	Fibrinogen	FG(A,B,G)	4q21–q31
Emphysema-cirrhosis	10740	α-1 Antitrypsin (protease inhibitor)	PI	14q24.3–q32.2
Hemophilia A	30670	Antihemophilic globulin A (factor VIII)	HEMA	Xq26–qter
Hemophilia B	30690	Factor IX	F9	Xq26–q28
Plasminogen activator deficiency*	17337	Plasminogen activator	PLA	6
Immune deficiency diseases				
C2 deficiency	21700	Complement C2	C2	6p21.3
C4 deficiency	12081	Complement C4	C4	6p21.3
C3 deficiency	12070	Complement C3	C3	19pter–q13.2
C8 deficiency*	12095,6	Complement C8	C8	1p
Immunodeficiency	19171	Uridine monophosphate kinase	UMPK	1p32
Nucleoside phosphorylase deficiency	16405	Nucleoside phosphorylase	NP	14q131
Severe combined immunodeficiency	10270	Adenosine deaminase	ADA	20q13.2–qter
Endocrinopathies				
Congenital adrenal hyperplasia	20191	21-Hydroxylase deficiency	CA21H	6p21.3
Growth hormone deficiency	13925	Growth hormone	GH1	17q22–q24
Hyperinsulinemia	17673	Insulin	INS	11p15
Diseases (biochemical defect unknown)				
Adrenoleukodystrophy	30010		ALD	Xq
Cataract, zonular pulverulent	11620		CAE	1cen–q25
Color-blindness (deutan)	30380		CBD	Xq27–qter
Color-blindness (protan)	30390		CBP	Xq27–qter
Chronic granulomatous disease	30640		CGD	Xpter–p21
Charcot-Marie-Tooth disease (slow conduction type)	11820		CMT1	1q
Dentinogenesis imperfecta 1	12549		DGI1	4
DiGeorge syndrome	18840		DGS	22q11
Hemochromatosis	23520		HFE	6p
Huntington's disease	14310		HD	4
Langer-Giedion syndrome	15023		LGS	8q23.3
Muscular dystrophy, Becker type	31010		BMD	X
Muscular dystrophy, Duchenne type	31020		DMD	Xp21
Myotonic dystrophy	16090		DM	19
Nail-patella syndrome	16120		NPS1	9q3
Ocular albinism	30050		OA	Xp22
Prader-Willi syndrome	17627		PWS	15q11
Retinoschisis	31270		RS	Xp
Sclerotylosis	18160		TYS	4
Spinocerebellar ataxia	16440		SCA1	6
von Recklinghausen neurofibromatosis 1	16220		NF1	19
Malignant disorders				
Retinoblastoma-1	18020		RB1	13q14
Wilms' tumor – aniridia/ambiguous genitalia/mental retardation	19407		WAGR	11p13
Miscellaneous				
Ehlers Danlos syndrome VII*	13006	Collagen I (α-2)	COL1A2	7q21–q22
Osteogenesis imperfecta* (one or more types)		Collagen I (α-2)	COL1A2	7q21–q22
Osteogenesis imperfecta (one or more types)		α-1 Type 1 collagen	COL1A1	17q21–22
X-Linked ichthyosis	30810	Steroid sulfatase	STS	Xpter–p22.3

---

* Provisional assignment at time of writing.

† Adapted from McKusick (1980b) and Shows, Sakaguchi and Naylor (1980) with recent updated information.

** *See* McKusick (1983).

Table 3.7 Hereditary disorders in which genetic linkage has clinical application

Disease	*McKusick[*] number*	*Linked marker*	*Chromosome*	*Indications*
Zonular pulverulent cataract	11620	Duffy	1q	Prenatal diagnosis
Elliptocytosis-1	13050	Rh	1p	Prenatal diagnosis
Charcot-Marie-Tooth disease-1	11820	Duffy	1q	Genetic counseling
Huntington's disease	14310	DNA polymorphism	4p(?)	Prenatal diagnosis, genetic counseling
Dentinogenesis imperfecta-1	12549	Group specific component	4q	Genetic counseling
Sclerotylosis	18160	MNSs blood group	4q	Genetic counseling
Congenital adrenal hyperplasia (21-hydroxylase deficiency)	20191	HLA	6p	Prenatal diagnosis, early therapy
Spinocerebellar ataxia-1	16440	HLA	6p	Prenatal counseling
Hemochromatosis	14160	HLA	6p	Early diagnosis and prophylactic therapy
C2 deficiency	21700	HLA	6p	Prenatal diagnosis
C4 deficiency	12981	HLA	6p	Prenatal diagnosis
Nail-patella syndrome	16120	ABO	9q	Prenatal diagnosis
β-Hemoglobinopathies		DNA polymorphism (β-gene probe)	11p	Prenatal diagnosis
Phenylketonuria	26160	DNA polymorphism	12(?)	Prenatal diagnosis, genetic counseling
Retinoblastoma	18020	Esterase D	13q	Intensive surveillance, prenatal diagnosis
Norum disease (LCAT deficiency)	24590	Haptoglobin	16q	Genetic counseling
Familial hypercholesterolemia	14389	C3, DNA polymorphism	19	Prenatal diagnosis
Myotonic dystrophy	16090	Secretor, Lutheran, peptidase D, C3, DNA polymorphism (C3)	19	Prenatal diagnosis, genetic counseling
Hemophilia A	30670	G6PD	Xq	Prenatal diagnosis
Duchenne muscular dystrophy	31020	DNA polymorphism	Xp	Prenatal diagnosis, genetic counseling

[*] *See* McKusick (1983).

Gene mapping information will also permit the detection of genetic heterogeneity of diseases, which clinically appear to be the same or similar but have a different genetic basis.

Gene mapping information also has considerable relevance to cytogenetics. Although there have been major advances in the development of differential banding and staining techniques, the current chromosome methodology still has major limitations. For example, the definition of exact break points in structural chromosomal rearrangements may be difficult and gene mapping information, through the study of genes known to be located around the break points, will help to pinpoint the exact part of the chromosome affected. Gene mapping information also may help to distinguish between balanced and unbalanced structural chromosomal rearrangements. Sometimes, extra chromosomal material may be too small in size to permit certain identification and the use of gene dosage studies of genes located on the extra chromosome material will permit confirmation of the suspected cytogenetic interpretation. Chromosome studies may take several hours or days to complete, but the use of gene dosage studies may permit more rapid diagnosis of a chromosome problem, or its confirmation. If one can identify from which parent a chromosome abnormality has originated, it may be possible, as causes of chromosome abnormalities are recognized, to utilize preventive measures in future pregnancies to reduce the likelihood of a chromosome abnormality recurring; the parental origin of the chromosome imbalance can be established through the use of polymorphic markers. Finally, if one can understand the basis of phenotypic abnormalities with a chromosome problem, it may be possible to institute some form of therapy, whose nature is not obvious at the present time. Insight into the basis of the phenotypic abnormality may come through knowing which genes are located on an affected chromosome.

FUTURE DEVELOPMENTS

With current techniques, especially those utilizing DNA strategies, it is now theoretically possible to map the complete human genome. While this will be a major undertaking, it is now technically feasible and the time needed to accomplish this will depend upon the commitment of effort and resources. With continued progress being made in the development of new techniques, it is conceivable that complete mapping of the human genome may be accomplished with less effort and time than appears to be the case at present. As the human map expands, it will be possible to apply it in an ever broadening number of problems, both at the clinical and basic biological level.

Glossary

Association: The occurrence together in a population or kindred of two or more phenotypes with a frequency greater than would be anticipated on the basis of chance alone. This term is sometimes used interchangeably with linkage, which refers to a different situation.

Chromatid: One of the two parallel arms of the prophase or metaphase chromosome. These arms are joined at the centromere and, at the time of cell division, each will be distributed to a different daughter cell.

Chromosome: This is the physical location of the nuclear genes. They are generally visible only at the time of cell division or just prior to this and have a characteristic number in a species, 46 chromosomes being typical for normal humans.

Co-dominance: The state when both alleles are expressed in the heterozygote.

Linkage: Genetic loci are linked when found close together on the same chromosome.

Morgan: The unit of genetic map distance between cross-over events.

Oncogene: Viral oncogenes occur in acutely transforming RNA tumor viruses and appear to be responsible for the transformation properties of these viruses. Homologous nuclear DNA sequences are found in many normal cells, including those of man. These are termed cellular oncogenes. Evidence indicates that the cellular oncogenes are the evolutionary progenitors of the viral oncogenes.

Recombination: The process by which genes on the chromosome end up on the homologous chromosome is called crossing-over. Recombination refers to the result of crossing-over.

Restriction fragment length polymorphism (RFLP): Different lengths of DNA from the same region resulting from treatment with a specific restriction endonuclease. The different lengths result from different base sequences at a given site on the DNA. The use of the RFLPs will greatly facilitate human gene mapping and linkage studies.

Synteny: Presence together on the same chromosome irrespective of the distance between genes.

References

BENEDICT, W. F., MURPHREE, A. L., BANERJEE, A., SPINNA, C. A., SPARKES, M. C. and SPARKES, R. S. (1983) Patient with 13 chromosome deletion: evidence that the retinoblastoma gene is a recessive cancer gene. *Science*, 219, 973–975

BOTSTEIN, D., WHITE, R. L., SKOLNICK, M. and DAVIS, R. W. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, 32, 314–331

CAVENEE, W. K., DRYJA, T. P., PHILLIPS, R. A. *et al.* (1983) Expression of recessive alleles by chromosomal mechanisms in retinoblastoma. *Nature*, 305, 779–784

CONNEALLY, P. M. and RIVAS, M. L. (1980) Linkage analysis in man. In *Advances in Human Genetics*, edited by H. Harris and K. Hirschhorn, Vol. 10, pp. 209–266. New York: Plenum Press

DE LA CHAPELLE, A. and BERGER, R. (1984) Report of the committee on chromosome rearrangements in neoplasia and on fragile sites. *Cytogenetics and Cell Genetics*, 37, 274–311

GODBOUT, R., DRYJA, T. P., SQUIRE, J., GALLIE, B. C. and PHILLIPS, R. A. (1983) Somatic inactivation of genes on chromosome 13 is a common event in retinoblastoma. *Nature*, 304, 451–453

GUSELLA, J. F., WEXLER, N. S., CONNEALLY, P. M. *et al.* (1983) A polymorphic DNA marker genetically linked to Huntington's disease. *Nature*, 306, 234–238

HODGE, S. E. and SPENCE, M. A. (1981) Some epistatic two locus models of disease. II. The confounding of linkage and association. *American Journal of Human Genetics*, 33, 396–406

HUMAN GENE MAPPING VII (1984) Seventh international workshop on human gene mapping. *Cytogenetics and Cell Genetics*, 37, 1–666

KAN, Y. W. and DOZY, A. M. (1978) Antenatal diagnosis of sickle-cell anemia by DNA analysis of amniotic fluid cells. *Lancet*, 2, 910–911

KAO, F. T. (1983) Somatic cell genetics cell genetics and gene mapping. *International Review of Cytology*, 85, 109–146

LANGE, K. and BOEHNKE, M. (1982) How many polymorphic genes will it take to span the human genome? *American Journal of Human Genetics*, 34, 842–845

LEBO, R. (1982) Chromosome sorting and DNA sequence localization. *Cytometry*, 3, 145–154

MCKUSICK, V. A. (1980a) The anatomy of the human genome. *Journal of Heredity*, 71, 370–391

MCKUSICK, V. A. (1980b) The anatomy of the human genome. *American Journal of Medicine*, 69, 267–276

MCKUSICK, V. A. (1982) The human genome through the eyes of a clinical geneticist. *Cytogenetics and Cell Genetics*, 32, 7–23

MCKUSICK, V. A. (1983) *Mendelian Inheritance in Man. Catalogs of Autosomal Dominant, Autosomal Recessive, and X-Linked Phenotypes*, 6th edn. Baltimore: The Johns Hopkins University Press

MCKUSICK, V. A. and RUDDLE, F. H. (1977) The status of the gene map of the human chromosomes. *Science*, 196, 390–405

MITELMAN, F. (1983) Catalogue of chromosome aberrations in cancer. *Cytogenetics and Cell Genetics*, 36, 1–516

MOHR, J. (1951) Estimation of linkage between the Lutheran and the Lewis blood groups. *Acta Pathologica et Microbiologica Scandinavica*, 29, 339–344

MORTON, N. E. (1955) Sequential tests for the detection of linkage. *Human Genetics*, 7, 277–318

OTT, J. (1974) Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. *American Journal of Human Genetics*, 12, 119–124

OTT, J. (1977) Linkage analysis with misclassification at one locus. *Clinical Genetics*, 12, 119–124

POLLACK, M. S., LEVINE, L. S., PANG, S. *et al.* (1979) Prenatal diagnosis of congenital adrenal hyperplasia (21-hydroxylase deficiency) by HLA typing. *Lancet*, 1, 1107–1108

RENWICK, J. H. (1971) The mapping of human chromosomes. In *Annual Review of Genetics*, edited by H. L. Roman, L. M. Sandler and A. Campbell, Vol. 5, pp. 81–120. Palo Alto: Annual Reviews Inc.

RUDDLE, F. H. (1981) A new era in mammalian gene mapping: somatic cell genetics and recombinant DNA methodologies. *Nature*, 294, 115–120

SCHROTT, H. G., KARP, L. and OMENN, G. S. (1973) Prenatal prediction in myotonic dystrophy: guidelines for genetic counseling. *Clinical Genetics*, 4, 38–45

SHOWS, T. B., SAKAGUCHI, A. Y. and NAYLOR, S. L. (1980) Mapping the human genome, cloned genes, DNA polymorphisms, and inherited disease. In *Advances in Human Genetics*, edited by H. Harris and K. Hirschhorn, Vol 12, pp. 341–452. New York: Plenum Press

SINET, P. M., ALLARD, D., LEJEUNE, J. and JEROME, H. (1974) Augmentation d'activite de la superoxide dismutase erythrocytaire dans la trisomie pour le chromosome 21. *Comptes Rendus de l'Academie des Sciences*, 278, 3267–3270

SINISCALCO, M. (1979) Approaches to human linkage. In *Progress in Medical Genetics (New Series)*, edited by A. G. Steinberg, A. G. Bearn, A. G. Motulsky and B. Childs, Vol. 3, pp. 221–307. Philadelphia: W. B. Saunders

SPARKES, R. S., MURPHREE, A. L., LINGUA, R. W. *et al.* (1983) Gene for hereditary retinoblastoma assigned to human chromosome 13 by linkage to esterase D. *Science*, 219, 971–973

SPARKES, R. S., SPARKES, M. C., WILSON, M. G. *et al.* (1980) Regional assignment of genes for human esterase D and retinoblastoma to chromosome band 13q14. *Science*, 208, 1042–1044

VOGEL, F. (1979) Genetics of retinoblastoma. *Human Genetics*, 52, 1–54

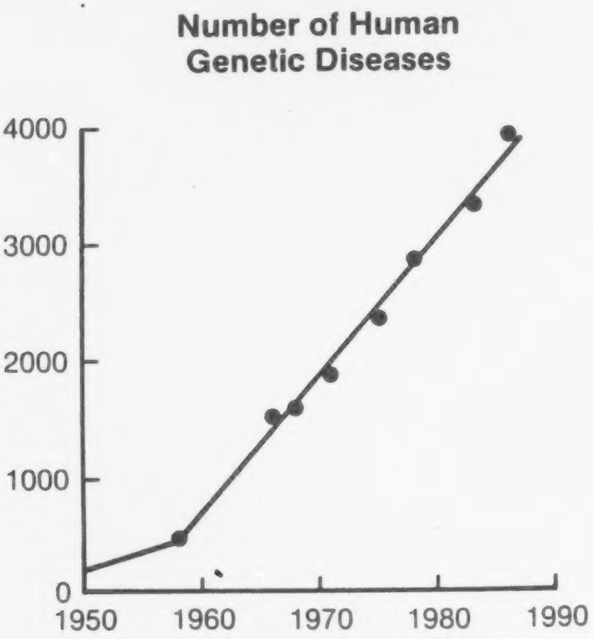
VOGEL, F. and MOTULSKY, A. G. (1979) *Human Genetics, Problems and Approaches*. New York: Springer-Verlag

WILSON, E. B. (1911) The sex chromosomes. *Archiv for Mikroskop Anatomie*, 77, 249–271

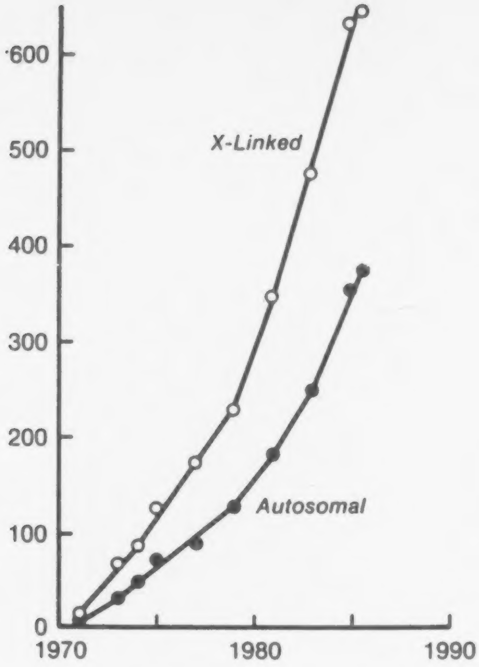
WOO, S. L. C., LIDSKY, A. S., GUTTLER, F., CHANDRA, T. and ROBSON, K. J. H. (1983) Cloned human phenylalanine hydroxylase gene allows prenatal diagnosis and carrier detection of classical phenylketonuria. *Nature*, 306, 151–155

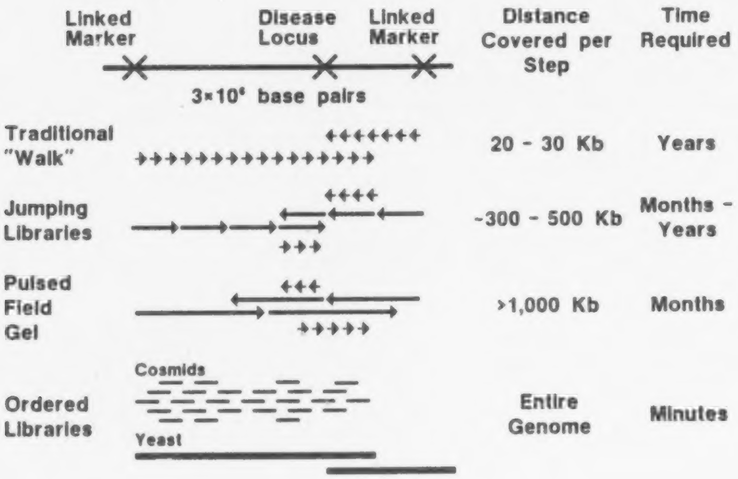
MEDICAL IMPLICATIONS
OF EXTENSIVE PHYSICAL AND SEQUENCE CHARACTERIZATION
OF THE HUMAN GENOME

Theodore Friedmann, M.D.
Department of Pediatrics and
Center for Molecular Genetics UCSD
School of Medicine
La Jolla, CA 92093


Number of Human Genetic Diseases
4000
3000
2000
1000
0
1950
1960
1970
1980
1990

Number of Mapped Genetic Loci on the Human Genome





Linked Marker
Disease Locus
Linked Marker
3×10⁶ base pairs
Cosmids
Yeast

	Distance Covered per Step	Time Required
Traditional "Walk"	20 - 30 Kb	Years
Jumping Libraries	~300 - 500 Kb	Months - Years
Pulsed Field Gel	>1,000 Kb	Months
Ordered Libraries	Entire Genome	Minutes

SUMMARY

Large-scale studies to characterize the human genome by physical mapping and nucleotide sequencing methods will soon be undertaken. The scientific rationalization for this "human genome initiative" includes the promise that isolation and understanding of many disease-related loci will be facilitated and that such information will lead to improved methods of disease detection, prevention and treatment. The approach involving the preparation of a fine-structure physical map and an overlapping contiguous set of clones spanning entire chromosomes or large portions of chromosomes promises to have an important and rapid impact on the management of human genetic disease.

During the past several years, methods have become available that promise to allow the entire human genome to be completely defined at the nucleotide level. It is a contentious issue, one that involves questions of scientific and medical priorities and goals, allocation of funding and other scientific resources and the relative roles of government and the academic and business communities in the pursuit of science and the delivery of health care. The direct and implied goals of much of modern biomedical research are to develop an understanding of the causes and mechanisns of pathogenesis of human disease, to devise effective diagnostic and screening procedures for early intervention, to develop effective methods of prevention, to develop truly effective methods of therapy and to bring all of these within reach of all who need care. It is widely assumed that a complete characterization of the human genome - the so-called "human genome initiative" - will lead to major improvements in the understanding and management of many forms of human disease, and it is the purpose of this report is to examine that assumption, to review the present approaches to the characterization and management of human genetic diseases and to examine the potential impact that a large-scale physical and genetic characterization of the human genome may have on the delivery and the effectiveness of future medical care. To do so and to understand some of the medical pressures that affect attitudes toward a large-scale human genome characterization, it is helpful to know the state of effectiveness of current approaches to genetic disease management.

A METABOLIC APPROACH TO GENETIC DISEASE

It is very important to emphasize that the term genetic disease now clearly describes not only the more or less classical, and individually rather rare disorders showing Mendelian inheritance, but also many extremely common diseases demonstrating complicated multifactorial and possibly multigenic inheritance such as the immensely important cardiovascular disorders, diabetes mellitus and neuropsychiatric diseases, and even defects resulting from somatic cell mutations, such as probably most of human cancer.

The present approach toward the detection, control and therapy of human genetic disorders has its conceptual origin in the re-discovery at the beginning of the 20th century of the laws of Mendel and their application to human inherited disorders by the British physician Sir Archibald Garrod.[1] Garrod developed our current concept that some human genetic diseases result from a deficiency of enzymes and that such deficiencies cause "inborn errors of metabolism" by interfering with the production of required metabolites or by causing an accumulation of toxic precursors prior to the metabolic block. The development of the sciences of biochemistry, genetics and cell biology during the ensuing first half of the 20th century and the birth and maturation of molecular biology led to sophisticated knowledge of the genetic mechanisms responsible for a large number of human disorders. The role of DNA as the repository of genetic information was established, the nature and mechanisms of mutations came to be characterized, the flow of normal and mutant genetic information from DNA to messenger RNA to functional or defective gene products - the proteins - became understood and the mechanisms of normal and mutant protein structure-function relationships in disease were becoming clarified.

These advances have made it possible to identify the enzymic or other protein defects in many genetic diseases and to develop assays for the involved metabolic enzymes or for substrates or metabolic products produced and stored in cells as a result of enzymic defects. Because a number of genetic diseases are reflected in abnormal gene expression in cultured fibroblasts or other cells it has become possible to study those disorders much more effectively in vitro than in intact patients or in their tissues, and to develop the powerful and conceptually new approaches of prenatal diagnosis based on a demonstration of abnormal gene expression in fetus derived amniotic or chorionic villus cells.[2,3]

Since the demonstration of the amino acid sequence alteration in sickle cell hemoglobin by Pauling and his colleagues in 1949[4] and the demonstration of the enzyme defects beginning with galactosemia,[5] phenylketonuria,[6] von Gierke's disease[7] and others in the 1950s and 1960s, a large number of human genetic diseases have been defined and their enzyme or other protein defects identified. By 1958, there were 412 recognized genetic disorders, and in 103 of these the enzyme defects were understood (Fig. 1 and Table 1).[8] Since then, that number has grown to a total of almost 4,000 recognized genetic loci or disorders, in more than 200 of which the enzymic defect has been demonstrated.[9] In other diseases, such as the hemoglobinopathies, deficiencies of coagulation factors and other circulating proteins, disorders of collagen structure, diseases of membrane transport, deficiencies of binding proteins, receptors or peptide hormones and many others, defects in non-enzyme proteins have also been demonstrated.

Unfortunately, this now classical metabolic approach to genetic disease - i.e., the detection of the abnormal enzyme or its aberrant - is hampered by some severe limitations not immediately obvious in light of the rather impressive growth in the number of characterized human diseases. In order to characterize and manage diseases by this approach, it is necessary to demonstrate and understand the biochemical defect, a task that is still extremely difficult and beyond reach for many or even most genetic diseases. Often there are few or no clues to the nature of the abnormal metabolic step, as in the case of cystic fibrosis.[10] Without an assay for the defective enzyme or metabolite, accurate and reliable diagnosis and therapy are difficult or impossible. In such cases, the clarification of genetic disease therefore becomes a slow, tedious, usually disappointing search for the biochemical defect. Furthermore, many disease-related genes may be expressed only in one or few cell types or tissues or in relatively inaccessible specific organs such as the liver or brain, making diagnosis difficult. For some disorders, the expression of mutant genes may lead quickly and very early in life to irreversible damage, and the need to await inevitable physiological or cellular damage and the establishment of overt clinical disease before diagnosis or genetic counselling start may destine patients to an unfavorable outcome. And certainly, the task of identifying the combined and interactive biochemical and environmental factors in complex multigenic diseases is even more daunting.

Equally troublesome is the realization that the application of biochemical and metabolic understanding of genetic disease to achieve truly effective therapy has only rarely been successful. While there have been some striking therapeutic successes, such as in dietary therapy for phenylketonuria and galactosemia, enzyme co-factor therapy for methylmalonic acidemia and multiple carboxylase deficiency, gene product replacement for forms of insulin-dependent diabetes mellitus, hemophilia and thyroid and other hormone deficiencies, pharmacological therapy aimed at the removal of toxic stored products as in Wilson's disease and even tissue transplantation for forms of thalassemia and for a variety of storage diseases such as Hurler's and Maroteaux-Lamy syndromes, it is fair to say that most genetic diseases cannot yet be well treated by manipulations of the relevant metabolic pathways.[11-15]

DIAGNOSIS BY DETECTION OF MUTANT GENES

It seems self evident that a more detailed understanding of diseases-related genes, of their structure and of their mechanisms of regulated expression will lead to improved and more effective diagnostic and therapeutic techniques. The management of human disease through the detection of the mutant genes rather than through the metabolic activities of expressed mutant gene products offers, in principle, several advantages;

1. a more complete understanding of the pathogenesis of disease through a study of gene structure and function.

2. simplified diagnostic and screening procedures. Since any source of cells or DNA would be suitable and because gene expression is not required for detection, difficulties imposed by tissue or organ-specific expression would be reduced or even eliminated.

3. detection of mutant genes may be carried out before cell or tissue damage takes place, and therefore accurate diagnosis and effective treatment would be independent of, and precede active disease.

"Forward Genetics"

Through the development of the tools of molecular genetics during the past several decades, great progress has been made in understanding and characterizing the role of mutant genes in the almost 4000 human genetic diseases. Until recently, most methods for cloning specific target genes have required prior characterization of the biochemical defect responsible for the disease. This direct genetic approach requires the identification and isolation of the mutant gene product followed by the determination of at least partial amino acid sequence of the defective protein, the development of specific antibodies, the use of in vitro selection conditions for cells expressing or failing to express the protein or the availability of highly differentiated cells that produce very large amounts of a specific protein. With a gene product available in pure form, it becomes possible to devise cloning strategies based on the use of synthetic oligonucleotides or antibodies to screen several different kinds of gene libraries - both genomic and cDNA libraries. Cloning strategies based on this scheme are therefore useful and applicable for disorders in which the mutant enzyme or other protein can be purified and at least partially characterized. Such a traditional approach has come to be called "forward" genetics, although paradoxically it follows a reverse pathway along the scheme of gene expression from gene to messenger to protein gene products. The traditional forward genetic approach requires:

1. Knowledge that the disease phenotype results from a genetic factor. This is most often established by family studies.

2. Identification of the biochemical or metabolic defect.

3. Purification and and at least partial characterization of the mutant gene product.

4. Isolation and cloning of the normal and mutant genes and characterization of their function.

5. Use of the cloned, characterized gene sequences to develop molecular probes for diagnosis and possibly for

therapy.

The isolation and cloning of important human genes by this approach will obviously continue at the rate limited by the efficiency with which new genetic defects are biochemically characterized and the relevant proteins purified and characterized. Unfortunately, for most human genetic disorders, the responsible gene product or products have not yet been identified, and it is therefore not yet possible to characterize and understand the genes involved in most human genetic diseases by "forward" genetics. It is likely that extremely sensitive amino acid sequenators will become available in the relatively near future that will allow the determination of sufficient N-terminal amino acid sequence from individual spots on two-dimensional protein gels to permit the synthesis of oligonucleotide probes for gene cloning. Two-dimensional protein gels will be prepared from normal and diseased cells or tissues and the very small amounts of mutant and their counterpart normal proteins will be eluted from the gels and used for partial amino acid sequence determination and possibly for the preparation of antibodies for "forward" genetic analysis as described above. In this approach to disease, it is not necessary to know anything at all of the biochemical, structural or enzymic functions and properties of the disease- related gene. While technical advances in protein isolation and characterization may soon increase the efficiency of detection and characterization of mutant gene products, the identification of the defective gene product or products in many genetic diseases, especially the polygenic ones, will continue to be a major obstacle to the "forward" genetic approach.

Nonetheless, for a rapidly growing number of human disease-related genes, this forward, protein-based approach has already proven to be very powerful and useful, starting with the genes encoding the globin proteins,[16] followed by a large and rapidly growing number of other human disease-related genes.[17] Using molecular probes isolated by these procedures, it has been possible to devise and offer disease detection and screening programs for major, high-burden diseases such as sickle cell anemia and the thalassemias to appropriate target populations as well as individual and family counseling for many other genetic defects, although the genetic heterogeneity of most genetic disorders makes population-wide screening programs unfeasible and impractical. In the case of several disorders involving circulating gene products such as blood coagulation factors, growth hormone, human insulin, tissue plasminogen activator (TPA) and others, cloned genes have been used for the large-scale manufacture of the biologically active gene product for therapeutic application, and the use of many other cloned genes for the similar large-scale production of therapeutic reagents will obviously increase.

THE LINKAGE MAP OF THE HUMAN GENOME AND "REVERSE" GENETICS

There is obviously a great need to be able to identify and characterize genes associated with the vast majority of human genetic disorders with no identified biochemical defect and no known target gene products, i.e., diseases that are therefore not amenable to the "forward" genetic approach. The approach to this problem is conceptually the reverse of the traditional scheme, and involves first the isolation of a target gene or other gene-related sequence followed by the identification of the function of the sequence or the encoded gene product. Such an approach to the study or detection of a genetic disease has come to be called "reverse" genetics, and, as with "forward" genetics, aims to isolate disease-related loci, to characterize their structure and their relationship to the development of disease and to use isolated genes or portions of genes to develop diagnostic reagents and tools and to manufacture therapeutic materials. This approach uses as its starting point nucleotide sequences known to map at or near the locus for the disease followed by isolation of the entire relevant nucleotide sequence by gene mapping and chromosomal "walking" procedures.

Until several years ago, the "reverse" genetic approach might have seemed extremely cumbersome, but it has now become not only feasible but in fact one of the most powerful, effective and promising approaches to understanding the bulk of human genetic disease. This scheme depends not only on efficient cloning and chromosomal "walking" methods, but most crucially on the development of a thorough genetic, or linkage, map of the human genome. Such a linkage map is constructed by determining the frequency of exchange of genetic material or recombination between homologous chromosomes during meiosis. A traditional approach to the development of such a map would be to use classical genetic loci, i.e., those for recognized gene products, whose approximate genomic locations are known.

Human gene mapping was virtually non-existent until less than twenty years ago. In 1971, only 15 human markers had been localized to specific chromosomes (Fig. 2),[8] 12 of them to the X chromosome because of the ease of identifying X-linked disorders by family studies. There were only 11 recognized sets of linked loci, only 3 of which contained 3 loci and the remaining sets contained only two linked loci each. The human genome has been estimated to contain approximately $3x10^9$ nucleotide base pairs and to extend over approximately 3300 centimorgans, where a centimorgan represents a unit of genetic distance corresponding to 1% recombination. It is thought to correspond very approximately to an average of 10^6 nucleotide base pairs. Many hundreds of mapped markers would therefore be required to achieve a linkage map at even 10 centimorgan (cM) (10% recombination) resolution, a distance that begins to be useful for purposes of genetic counselling.

From 1971-1986, there was a rapid increase in the chromosomal assignments of genetic loci (Fig.2). By 1979, there were 123 confirmed autosomal and 230 confirmed X-linked markers available for linkage analysis. Beginning in 1978, it became evident that the human genome contains markers other than traditional gene loci that could be used for the determination of a genetic map of a complex genome such as that of the human. These genetic markers are the large number of normal variations in nucleotide sequence throughout the genome, some of which are located within restriction enzyme recognition sites and which therefore cause restriction fragment length polymorphisms (RLFPs) after digestion of genomic DNA with restriction enzymes. It of course became quickly evident that such RFLPs could be used just as any other genetic loci could be used, not only for diagnostic purposes but also for linkage analysis.[18,19] Partly as a result of the use of RFLP markers, the rate of acquisition of new genetic mapping data increased markedly thereafter, to the point where a human genetic map almost saturated at the 3-5 centimorgan level is very close to becoming a reality. The entire human genome therefore now has milestones scattered throughout virtually its entire length at distances amenable to physical characterization by currently available or imminent methods.

The use of linkage analysis, using RFLPs or other markers physically separate from a target disease locus but sufficiently close to be detectably linked to it, provides the opportunity to identify the map location of any disease locus, even in the cases where there is absolutely no information on the nature of the biochemical or genetic defect, and to search for and identify RFLPs that are linked tightly enough to the disease locus to be useful diagnostically. The recent striking successes with the development of such useful RFLP markers for previously inaccessible diseases such as Huntington's disease,[20] cystic fibrosis,[21] Duchenne muscular dystrophy[22] and the beginnings of similar studies for colonic polyposis,[23] neurofibromatosis,[24] Alzheimer's disease,[25] bipolar disease[26] and many others, indicates the importance of a saturated human genetic linkage map for the development of diagnostic reagents for all of the other human genetic diseases.

Not only has such an approach proved clinically important for preparing diagnostic materials for otherwise uncharacterized genetic diseases, but sequences revealed by RFLPs that are linked to a disease provide a starting point for the isolation of the relevant disease locus itself by a variety of physical methods including genomic "walking", physical mapping with pulsed-field gel eletrophoresis and other, direct techniques to isolate the sequences of the target gene itself. This step in "reverse" genetics - i.e., the isolation of the specific disease-related gene rather than diagnostically useful but distant markers - would permit the identification of the relevant gene product and an educated guess of its function through structural and evolutionary studies. The first striking successes using this approach have come from the isolation of the genes for chronic granulomatous disease[27] (CGD) and Duchenne muscular dystrophy.

The process in these cases has involved combinations of the following steps:

1. the identification of the approximate genomic location of the disease locus. Through the power of modern linkage analysis and the very imminent development of a complete human genetic linkage map, this step is or will soon be feasible for all relatively simple Mendelian human genetic diseases.

2. the use of nearby linked sequences to isolate sequences within the target gene. One possible approach has been the step-wise procedure of genomic "walking", a procedure reviewed extensively elsewhere.[28] Conceptually elegant modifications and variations on this general walking approach utilizing "linking" and "jumping" libraries[29,30] and efficient cosmid and other vectors have been developed by a number of investigators, but the unfortunate fact is that the methods of physical sequential translocation along a genome at the present time are extremely laborious and tedious and cannot be used easily to span the very enormous distances often required in most walking projects from a linked marker to a disease locus. For instance, a chromosomal walk of several hundred contiguous kilobases (1 kilobase = 1,000 bases) is a difficult and time consuming effort and has been

achieved at great effort only in a small number of laboratories. When one considers the fact that an average human chromosome, such as the X chromosome, may contain several hundred thousand kilobases and the distance between its 600 known loci would therefore average several thousand kilobases, it becomes obvious that there is a pressing need to develop methods to move much more rapidly from one gene locus to another on a chromosome. Technical advances are being made that will allow more efficient movement along the genome.

Other methods have been used to identify and isolate sequences from within the target gene without having to resort to tedious and inefficient sequential walking. Particularly helpful have been somatic cell hybrids or cells with deletions or other rearrangements in which the relative locations of genetic loci could be determined. The isolation of the Duchenne muscular dystrophy and CGD loci have taken advantage of cells with such well defined breakpoints and deletions.

3. demonstration that the isolated sequences encode true, expressed genes. Functional expressed genes have been identified by the use of candidate gene-specific DNA sequences to demonstrate transcribed messenger RNAs by screening cDNA libraries, by Northern blotting methods or by the demonstration by restriction mapping studies

of the presence of clustered, unmethylated CpG-rich "HTF" islands thought at least in some cases to be a "signature" of expressed genes.[31] Further proof of the importance of isolated, putative disease-related genes or portions of genes comes from a demonstration that probes for such sequences also detect the presumably conserved sequences in the genomes of a wide variety of non-human species, i.e., "zoo blots".

4. demonstration that the isolated sequences are expressed abnormally in cells from patients with the target disease. This crucial step is required to distinguish between the specific disease locus and other nearby but unrelated genes.

5. full sequence characterization of the isolated cDNA and genomic sequences from both normal and patient material in order to identify the nature of the abnormal gene product, its metabolic or structural role and to understand mechanisms of pathogenesis.

While these are impressive and useful techniques, they are difficult, laborious and time consuming. In the case of chronic granulomatous disease (CGD),[27] the "reverse" genetic isolation and characterization of the gene beginning with mapping data and proceeding to the gene itself was relatively rapid. But, perhaps more indicative of the potential problems surely to be encountered frequently in similar "reverse" genetic projects have been the problems associated with the isolation of the Huntington's disease locus. It has been more than 4 years since the first published report of linkage of the RFLP marker G8 to Huntington's disease,[20] and even with the extensive, highly collaborative efforts of several large groups spanning probably hundreds of person years of work, the HD gene has remained elusive and has not been isolated. Of course, the availability of a more thoroughly saturated linkage map and the growth of this field of science now more densely populated with investigators, will speed progress in disease diagnosis and screening through the identification of important disease linkages to RFLPs and to the recently described, more highly informative variable number tandem repeat (VNTR) probes.[32] Since a linkage map of the human genome with a resolution of no more than 3-5 centimorgans is nearing completion, the rate with which genetic diseases will be mapped will now increase more rapidly than ever before. Immediately after the advent of RFLP linkage analysis for human disease in 1978, the rate with which genetic diseases became mapped increased approximately four-fold, reflecting of course not only the power of the techniques but also the increased interest in human genetic and other scientific communities to use the new methods. Nevertheless, the very difficult task of moving from the linked marker to the target disease locus itself by physical methods in order to characterize the gene and the disease precisely and to understand mechanisms of pathogenesis and probably of treatment, will continue to present major technical problems.

A PHYSICAL MAP OF THE HUMAN GENOME - APPROACHES AND MEDICAL IMPLICATIONS

A detailed genetic linkage map of the human genome will define the relative positions of all known genetic loci, allows the identification of RFLPs linked to most and perhaps to all human genetic diseases, and provides a starting position for genomic walks to the relevant disease loci themselves. On the other hand, a linkage map will not provide very much useful information on the vast intergenic regions of the genome that may not be reflected in human diseases, and will not solve the difficulties of walking the several thousand kilobase distances that will separate most markers on the genetic map. It is these problems that require a different solution, and for which the physical characterization of the human genome and the development of a fine-structure physical map holds so much promise.

A physical map will be similar, but not identical to, the genetic linkage map. It is known that there are discrepancies between the two kinds of map in the human, and that genetic distances defined by recombinational events do not always reflect the physical distance between markers. There are recombinational "hot spots" where genetic distance will give an overestimate of the physical distance, and there are recombinational "cold spots". Recombination varies between males and females and with position on the chromosome. Therefore, a linkage map represents a view of a genome through a distorting lens. For some disorders, such as the Fragile X syndrome,[33] the disease locus lies close to, and may indeed include, a recombinational hot spot, and the task of designing a walk from closely linked markers is made difficult by the problem of not knowing how far to walk and how to recognize the locus when it has been reached.

The ultimate goal of the development of a detailed physical map of the human genome is to obtain the ultimate map possible,i.e., the complete sequence of the human genome. The path toward this goal can be envisioned to take 3 directions, each with different implications for disease detection, characterization and treatment:

1. an immediate start on a total sequence determination of the genome, without any prior or weighted selection of target regions. This approach may utilize not only present rapid sequencing methods but also imminent developments or "order of magnitude" improvements in sequencing technology, including both manual and automated methods. Nevertheless, some workers consider present manual methods efficient enough to warrant a start at random sequence analysis of the entire human genome if a collaborative or highly organized sequencing effort is made.

2. the development of a fine-structure physical map of cloned, overlapping sequences to span the entire genome, chromosome by chromosome. This approach is an extension of currently available and functioning techniques already used successfully to study the genomes of E.coli and C.elegans. The entire E-coli genome has been ordered and aligned as a contiguous set of clones[34] and approximately 85% of the C.elegans genome has been ordered into a large number of contigs.[35] The C.elegans genome of approximately 80 X 10^6 nucleotide base pairs is approximately the same size as a small human chromosome such as #21 or a fragment of a larger chromosome, implying that such an approach may therefore already be technically feasible for a large region of the human genome. The recent development of yeast cloning vectors of higher cloning capacity than the previously used cosmid vectors[36] and the use of pulsed field electrophoretic mapping methods[37] makes this approach even more attractive. The most important feature of this approach, however, is the fact that, not only would it provide map information, but it would also make immediately available a new kind of reagent - a set of overlapping, cloned and ordered human sequences spanning an entire chromosome or portion of a chromosome, whose relationship to any and all disease loci on that chromosome or fragment of a chromosome would be known. It would therefore, in principle, not be necessary to do

difficult and tedious genomic walking to move from a linked RFLP to a disease locus. The walk would not not be down the genome, but rather to the freezer containing the ordered clones.

3. the development of a physical map, as above, combined with nucleotide sequence determination of selected, "important" or "interesting" large contigs containing disease loci of particular interest.

For most well understood human diseases, the isolation and characterization of genes has had a far greater impact on detection, diagnosis, screening and genetic counselling than on therapy. It is often tempting for scientists and medical investigators to overemphasize the speed with which detailed structural knowledge, including complete nucleotide sequence information, might become translated into efficient disease treatment. There are few if any genes whose structure and function are understood as well as those of the globin proteins, and, in fact, few diseases whose pathogenesis is as well understood as that for the hemoglobinopathies. And yet, progress toward really effective therapy for these diseases has not moved with nearly the same pace as biochemical and molecular understanding. Therefore, in all likelihood, knowledge gained from a detailed understanding of the structure and sequence of the human genome will continue to have its most rapid and immediate effect on detection, diagnosis, screening and genetic with longer range effects on therapy, whether it be gene more efficient forms of traditional therapy or conceptually new forms of gene therapy.

1. Direct, untargeted nucleotide sequencing.

The approach involving the immediate, random, non-targeted large-scale nucleotide sequence determination of the entire genome is the least likely to have major, immediate effects on the management of human disease but is certain to become increasingly feasible and assume a high priority in the longer term. This approach represents the ultimate exercise in "reverse" genetics - to define the function of vast stretches of nucleotide sequence from sequence alone and to map and understand all genetic functions. That is a very formidable task, given the present power of sequence analysis and the developing ability to predict present mechanisms of gene expression from sequence alone. The barrier is not so much the problem of generating the nucleotide sequence per se - which is likely to be feasible and perhaps even accomplished within several decades - as much as it will be relating sequence to function and then to dysfunction of human disease. Problems raised by the interrupted, overlapping and fluid nature of the genetic material, such as with size and extent of introns, alternative splicing, exon shuffling, efficient paradigms for distinguishing small but expressed open reading frames from non-functional open reading frames will make sequence interpretation difficult but certainly amenable to improved analytic methods eventually to be developed. Nevertheless, these problems aside, a sequencing project per se will - slowly at first and then with increasing efficiency - isolate, map and characterize gene and inter-gene sequences useful as direct or linked probes for disease detection and possibly disease treatment. Direct, untargetted sequencing will assume greater and greater scientific credibility in the near future but will lead to clinically useful reagents only after it is well under way. It is obvious that extensive interpretations of sequence information will require a great deal of sequence comparison, not only between individual humans to distinguish between polymorphisms and disease-related mutations, but also between species to identify the significance of conserved sequences.

2. Ordered library construction and physical mapping.

For the immediate present and the near future, the goal of preparing a detailed physical map of individual, whole human chromosomes isolated, for instance by sophisticated physical tools such as the fluorescence activated sorting methods of flow cytometry, or of large fragments of human chromosomes obtained from somatic cell hybrids, is not only the most feasible technically, but also promises the quickest route to clinically useful reagents.

An essential part of the technology involves the preparation of libraries of contiguous, overlapping sets of clones. As overlapping contiguous sets of clones become available for large regions of the genome, disease loci will become physically available almost as quickly, since the most difficult and rate-limiting step - genomic walking - can be eliminated (Fig.3). If ordered libraries had been available when the RFLP-linked probes were first identified for cystic fibrosis, Duchenne's muscular dystrophy, Fragile-X syndrome, Huntington's Disease, Alzheimer's disease, bipolar disease, all of those genes would probably already be in hand and available for studies of pathogenesis and for diagnosis, detection and effective genetic counseling. The same is obviously true of the hundreds of RFLP-linked disease loci that will be identified in the coming years.

Obviously, the isolation of specific regions of target genes or other loci will be followed by nucleotide sequence determination. Since such regions are likely to be reasonably small, there seems to be no barrier to rapid sequence characterization of many important genes with resulting improved understanding of their role if disease. Just as obvious, however, is the fact that some loci, such as that for Duchenne's muscular dystrophy, will be very large, in the range of several megabases, and would therefore present slightly more formidable sequencing undertakings.

3. Mapping and ordered library preparation with selective targeted nucleotide sequence determination.

Not only will a mapping and ordered library approach generate specific target genes and other regions requiring sequence determination, but, very soon after starting, sizeable contiguous and overlapping sets of cosmid or other clones (contigs) would be generated. The experience with the mapping projects of both E. coli and C. elegans has indicated that the generation of many contigs of several hundred kilobases in length is rapid, and sequences of this length now represent quite reasonable targets, certainly in collaborative settings for nucleotide sequencing, not only by the automated sequencing methods anticipated in the next several years, but even now by current manual methods and their modifications. These methods are within reach of most groups seriously committed to genomic characterization.

The goal of isolating and characterizing sequences that will illuminate the mechanisms of human genetic disease and provide diagnostic reagents is therefore likely to be achieved most efficiently by large-scale genetic mapping accompanied by the preparation of overlapping, contiguous libraries. Through such an approach, disease loci will be isolated and characterized and used for the preparation of diagnostic probes and for the synthesis of pharmacologically useful gene products. In most cases, investigators will use sequence determination to study specific questions of mechanisms of genetic disease.

IMPLICATION OF GENOME CHARACTERIZATION FOR CLINICAL MEDICINE

As with most other medical and scientific advances, genetic knowledge, techniques and reagents derived from an extensive characterization of the human genome will provide benefits but will come with costs - financial, scientific and societal. The obvious benefit is the eventual large scale control of a great deal of human disease, largely through effective screening programs, early disease detection, more accurate and effective genetic counseling and the development of effective therapy based on improved understanding of disease. The development of conceptually novel forms of therapy, including gene therapy, will also be accelerated by knowledge derived from mapping and sequence characterization of the human genome, but without doubt the major effect at least at the outset will be on the development of effective new programs for the detection and accurate prediction of genetic disease.

Problems raised by the advent of so much new knowledge of genetic diseases and of so many new screening and detection programs are the same problems already faced by modern Medicine in an era of increasingly limited resources, will, imagination and commitment, but magnified many fold. These include:

1. the choice of diseases for which large-scale detection and screening programs should be established. No doubt, if and when effective detection methods become available for major, common, high-burden diseases such as most forms of cancer, neuropsychiatric disorders, coronary artery disease, diabetes mellitus, they will and should represent some of the first target diseases. However, the management of diseases, no matter how common or severe, for which there is no effective therapy, raises many difficulties for patients and the health systems alike, as is clearly illustrated in the case of Huntington's disease. To what use shall we put the immense amount of genetic information to be obtained in cases where no effective therapy exists? Many severe but rare diseases will become susceptible to detection and control but will become "orphaned" or ignored

because of their low societal impact. Lesch-Nyhan disease and ADA-deficiency immune defects would be examples of such disorders. This too is not a new phenomenon in Medicine.

2. the choice of target populations. For some common, high-burden diseases such as coronary artery disease, hypercholesterolemia, cystic fibrosis, Fragile X syndrome, bipolar disease, Alzheimer's disease and others, large-scale population screening would obviously be important, but could probably not be carried out until gene-specific probes become available to replace RFLP analysis, since very extensive screening based on RFLP analysis is not feasible with present methods. Even then, the costs and logistical problems associated with such wide-spread programs will be enormous. There will be an opportunity to add many newly diagnosable diseases to those for which prenatal diagnosis is feasible. Which diseases? Are there indications for mandatory screening and detection? What effect would such wide-spread detection capability have on the troublesome controversy over abortion and on the already low quality discourse on this issue?

3. the financial cost of screening programs. With the likelihood that genetic diagnosis will be feasible for the most common and severe diseases of our society, the

overall costs of large-scale screening programs can become very great very quickly. In many instances, insurance companies and other third party carriers do not uniformly reimburse patients fully for heterozygote detection procedures or for genetic counseling, activities that will certainly increase markedly with the development of many new genetic diagnostic programs.

4. the adequacy of current medical institutions, both public and private, non-profit and for-profit, to handle the anticipated heavy load of new screening and diagnostic procedures that may eventuall involve a very large proportion, even the majority, of the population.

5. the confidentiality of genetic diagnostic information. Again, there is little conceptually new here, except for the magnitude of the problem and therefore its potential for use and abuse. It represents potentially one of the most serious and difficult aspects of the future predictive power of molecular human genetics, and will frequently raise its head to complicate issues of genetic screening prior to employment and of suitability and eligibility for insurance coverage .

6. medicolegal issues. In such a rapidly developing and changing field, services will not be distributed uniformly, and inadequately trained or prepared health

> care providers will be faced with a flood of new techniques and procedures. It is inevitable that questions will eventually be raised of breached confidentiality, genetic damage, wrongful life and medical negligence. Very difficult issues of jurisprudence, including issues of competence, negligence, duty and malpractice culpability and medicolegal responsibility will certainly surface with far greater frequency than at present.

None of these issues is unique to new molecular genetic techniques, but the explosive increase in the development and use of genetic diagnostic tools brought about by increased knowledge of the human genome and its diseases will cause them to become important factors in the wide-spread application of molecular genetic diagnostic procedures.

SUMMARY AND CONCLUSIONS

Impressive progress has been made during the past several decades in the understanding the pathogenesis of human genetic disease. The tools of molecular biology have allowed to isolation of many disease-related genes by "forward" and a few by "reverse" genetics, and the imminent completion of a complete human genetic linkage map will accelerate the genetic characterization of many more genetic diseases. The major impacts of the molecular characterization of human genetic diseases have been;

1. to increase markedly the number of human diseases that we now recognize to have major genetic components, and to emphasize us the fact that genetic diseases do not represent, as once thought, rather rare medical curiosities with negligible societal impact, but rather constitute a wide spectrum of both rare and extremely common diseases responsible for an immense amount of suffering in all human societies.

2. to allow the development of powerful new approaches to diagnosis, detection, screening and even therapy of these disorders aimed directly at the mutant genes rather than at the gene products. This should eventually allow much more accurate and specific management of human genetic disease and the genetic factors in many human maladies.

The so-called "human genome initiative" will lead to even more effective and pervasive genetic diagnostic and therapeutic procedures. The establishment of a physical map based on the preparation of ordered libraries of overlapping, contiguous clones of human chromosomes is now feasible. Such a characterization will provide a large number of new diagnostic tools and reagents for many important genetic diseases. Within the coming decade, extensive nucleotide sequence analysis of the human genome will also begin to provide clinically useful reagents.

The wide-scale availability and use of these genetic techniques for a large number of human disorders may not raise any truly novel problems for the delivery of clinical care to patients, but will certainly exacerbate current problems in the present health care system, including issues of just and equitable identification of target populations, cost resource allocation, the choice of target diseases for major emphasis, the establishment of delivery systems for massive new screening and detection programs, the training and the regulation of health care providers knowledeable in genetics and the preparation and education of the system of jurisprudence for the inevitable challenges to it.

Fig. 1. Rate of identification of human genetic diseases, as listed in McKusick.[39]

Fig. 2. Number of human genetic diseases mapped to human antosomes (.—.) or to the X chromosome (o—o), as reported by the Human Gene Workshops. The increased rate of chromosomal map assignments between 1979 and 1980 corresponds to the first reported uses of RFLP linkage analysis by Kan and Dozy[19] and by Botstein and colleagues[20]. Data are derived from McKusick.[39]

Fig. 3. A comparison of physical strategies for isolation of a disease-related locus starting from flanking linked markers, together with the approximate time required for gene isolation. The method of traditional genomic walking relies only on sequential short steps (small arrows) of no more than 20-30 kb each in most cases, while the use of longer-range methods of jumping libraries and pulsed field gel methods also requires a

final series of short-distance steps after the more long-distance mapping (longer arrows) has been achieved. The availability of an ordered, contiguous set of clones makes all walking steps unnecessary.

TABLE I

RECOGNIZED HUMAN GENETIC DISEASES*

	1958	1966	1968	1971	1975	1978	1983	1986
AUTOSOMAL	285	269	344	415	583	736	934	1172
DOMINANT		(586)	(449)	(528)	(635)	(753)	(893)	(1029)
AUTOSOMAL	89	237	280	365	466	531	588	610
RECESSIVE		(294)	(349)	(418)	(481)	(596)	(710)	(810)
X-LINKED	38	68	68	86	93	107	115	124
		(51)	(55)	(64)	(78)	(98)	(128)	(162)
TOTAL	412	574	682	866	1142	1364	1637	1906
		(913)	(853)	(1010)	(1194)	(1447)	(1731)	(2001)
GRAND TOTAL	412	1487	1545	1876	2336	2811	3368	3907

*V.A. McKusick, The Human Gene Map Newsletter.[39]

Numbers in parentheses indicate loci not confirmed or validated.[39]

TABLE II

NUMBER OF LOCI MAPPED ON HUMAN CHROMOSOMES*

	AUTOSOMAL		X-LINKED	
	CONFIRMED	PROVISIONAL	CONFIRMED	PROVISIONAL
1971	3		12	
1973	31	28	64	88
1974	48	32	86	91
1975	72	46	125	95
1977	83	82	176	102
1979	123	87	230	112
1981	180	120	345	116
1983	247	161	488	118
1985	351	216	633	123
1986	358	273	730	130

*V.A. McKusick, The Human Gene Map Newsletter.[39]

REFERENCES

1. **Childs, B.** 1970. Sir Archibald Garrod's conception of chemical individuality: A Modern Appreciation. New England Journal of Medicine, **282**: 71-77.

2. **Milunsky, A.** 1986. Genetic Disorders in the Fetus. Plenum Publishing Corp., New York.

3. **Blakemore, K. and Mahoney.** 1986. Chorionic villus sampling in Genetic Disorders in the Fetus, ed. by A. Milunsky. Plenum Publishing Corp., New York.

4. **Pauling, L., Itano, H., Singer, S. and Wells, I.** 1949. Sickle cell anemia: A Molecular Disease. Science **110**: 543-546.

5. **Isselbacher, K., Anderson, E., Kurakashi, K. and Kalkar, H.** 1956. Congenital galactosemia, a single enzymatic block in galactose metabolism. Science **123**:635-636.

6. **Jervis, G.** 1953. Phenylpyruvic oligophrenia deficiency of phenylalanine-oxidizing system. Proc. Soc. Exp. Biol. Med. **82**:514-515.

7. **Cori, G. and Cori, C.**1952. Glucose-6-phosphatase of the liver in glycogen storage disease. J. Biol. Chem. 199:661-667.

8. **McKusick, V.** Mendelian inheritance in Man. Johns Hopkins University Press. 1971.

9. **McKusick, V.** 1986 in Mendelian Inheritance in Man, Johns Hopkins University Press, Baltimore, pp.xxxi-xxxviii.

10. **Talamo, R., Rosenstein, B. and Berninger, R.** 1983. Cystic fibrosis, in Metabolic Basis of Inherited Disease, ed. by J. Stanbury, J. Wyndgaarden, D. Frederickson, J. Goldstein and M. Brown. McGraw Hill, pp.270-286.

11. Metabolic Basis of Inherited Disease, ed. by J. Stanbury, J. Wyndgaarden, D. Frederickson, J. Goldstein and M. Brown. McGraw Hill, New York, 1983.

12. **Sweetman, L. and Nyhan, W.** 1986. Inheritable, Biotin-treatable disorders and associated phenomena. Ann. Rev. Nutrition 6: 317-343.

13. **Krivit, W., Pierpont, M., Ayaz, K., Tsai, N., Ramsay, M., Jersey, J., Weisdorf, S., Sibley, R., Snover, D.,**

McGovern, M., Schwartz, M. and Desnick, R. 1984. Bone marrow transplantations in the Maroteaux-Lamy Syndrome. New England J. Med. 311: 1606-1611.

14. **Hobbs, J., Hugh-Jones, K., Barret, A., Byron, N., Chambers, D., Henry, K., James, D., Lucas, C., Rogers, T., Benson, P., Tansley, L., Patrick, A., Mossman, J. and Young, P.** 1982. Reversal of clinical features of Hurler's disease and biochemical improvement after treatment by bone marrow transplantation. Lancet 2: 709-712.

15. **Thomas, E.D.** 1985. Marrow transplantation for non malignant disorders. New England J. Med. 312: 46-47.

16. **Maniatis, T., Jim, G. K., Efstratiadis, A. and Kafatos, F.** 1976. Amplification and characterization of beta-globin gene synthesized in vitro. Cell 8: 163-182.

17. **Proceedings of Human Gene Mapping** workshop 9, Paris, 1987. Clinical Genetics, in press.
18. **Kan, Y.W. and Dozy, A.M.** 1978. Polymorphisms of DNA sequence adjacent to human beta-globin structural gene: Relationship to sickle mutation. Proc. Nat Acad. Sci. USA 75: 5631-5.

19. **Botstein, D., White, R., Skolnick, N. and Davis, R.** 1980. Construction of a genetic linkage map in Man using restriction fragment length polymorphisms. Am. J. Hum. Gen. 32:314-331.

20. **Gusella, J., Wexler, N., Conneally, P., Naylor, S., Anderson, M., Tanzi, R., Watkins, P., Ottina, K., Wallace, M., Sakaguchi, A., Young, A., Shoulson, I., Bonilla, E. and Martin, J. A.** 1983. Polymorphic DNA marker genetically linked to Huntington's disease. Nature **306**: 234-238.

21. **Estivill, X., Farrall, M., Scambler, P., Bell, G., Hawley, K., Lench, N., Bates, G., Kruyer, H., Frederick, P., Stanier, P., Watson, E., Williamson, R. and Wainwright, B.** 1987. A candidate for the cystic fibrosis locus isolated by selection for methylation-free island. Nature **326**: 840-845.

22. **Koennig, M., Hoffman, E., Bertelson, C., Monaco, A., Feener, C. and Kunkel, L.** 1987. Complete cloning of the Duchenne Muscular Dystrophy (DMD) cDNA gene in normal and affected individuals. Cell **50**: 509-517.

23. **Bodmer, W., Bailey, C., Bodmer, J. et al.** 1987. Localization of the gene for familial adenomatous polyposis in chromosome 5. Nature **328**: 614-616.

23b. **Solomon, E., Voss, R., Hall, V. et al.** Chromosome 5 allele loss in human colorectal carcinomas. nature **328**: 616-618.

24. **Seizinger, B., Rouleau, G., Ozelius, J. et al.** 1987. Genetic linkage of von Recklinghausen neurofibromatosis to the nerve growth factor receptor gene. Cell **49**: 589-594.

25. **St. George-Hyslop, Tanzi, P., Polinsky, R., Haines, J., Nee, L., Watkins, C. et al.** 1987. The genetic defect causing familial Alzheimer's disease maps on chromosome 21. Gene **235**: 885-890.

26. **Egeland, J., Gerhard, D., Pauls, D., Sussex, J., Kidd, K., Allen, C., Hostetter, A. and Houseman, D.** 1987. Biopolar affective disorders linked to DNA markers on chromosome 11. Nature **325**: 783-787.

27. **Royer, B., Kunkel, L., Monaco, A., Goff, S., Newburger, P., Baehner, R., Cole, F., Curnutte, J. and Orkin, S.** 1987. Cloning the gene for an inherited human disorder - chronic granulomatous disease - on the basis of its chromosomal location. Nature **322**: 32-38.

28. **Carroll, M., Campbell, R., Bentley, D. and Porter, R.**

1984. Nature 307: 237-241.

29. **Collins, F. and Weissman, S.** 1984. Directional cloning of DNA fragments at a large distance from an initial probe: a circularization method. Proc. Nat. Acad. Sci. USA 81: 6812-6816.

30. **Poutska, A. and Lehrach, H.** 1986. Jumping libraries and linking libraries: the next generation of molecular tools in mammalian genetics. Trends in Genetics 2: 174-179.

31. **Bird, A.** 1986. CpG-rich islands and the function of DNA methylation. Nature 321: 209-213.

32. **Nakamura, Y., Leppert,M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Humlin, E. and White, R.** 1987. Variable number of tandem repeat (VNTR) markers forhuman gene mapping. Nature 235: 1616-1622.

33. **Nussbaum, R. and Ledbetter, D.** 1986. Fragile X syndrome: a unique mutation in man. Ann. Rev. Genet. 20: 109-145.

34. **Kohara, Y., Akiyama, K. and Isono, K.** 1987. The physical map of the whole E.coli chromosome:

application of a new strategy for rapid analysis and sorting of a large genomic library. Cell 50: 495-508.

35. **Coulson, A., Sulston, J., Brenner, S. and Karn, J.** 1986. Toward a physical map of the genome of the nematode C.elegans. Proc. Nat. Acad. Sci. USA 81: 6812-6816.

36. **Burke, D., Carle, G. and Olson, M.** 1987. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science 236: 806-812.

37. **Schwartz, D. and Cantor, C.** 1984. Separation of yeast chromosome-sized DNAs by pulsed field gel electrophoresis. Cell 37: 67-75.

38. **McKusick, V.** 1986. The Human Gene Map Newsletter, Oct. 15, 1986, p. 2.

MAPPING THE HUMAN GENOME: SOME IMPLICATIONS.

A REPORT TO THE OFFICE OF TECHNOLOGY ASSESSMENT OF THE CONGRESS OF THE UNITED STATES.

Jonathan Glover.

CONTENTS.

INTRODUCTION. 1

PART ONE: APPLICATIONS OF THE KNOWLEDGE: POSITIVE GENETIC ENGINEERING. 2
1. Some possible benefits. 4
2. Risks and mistakes. 5
3. Who would decide? 8
4. Justice and equality. 12
5. Autonomy and respect for persons. 14
6. Implications. 19

PART TWO: THE KNOWLEDGE ITSELF: ITS IMPACT ON ATTITUDES TO PEOPLE. 21

A. EQUALITY AND INEQUALITY OF GENETIC POTENTIAL. 21
1. The impact of discovering that we are genetically more equal. 22
2. The impact of discovering that we are genetically more unequal. 24
    (a) Allocation of resources. 24
    (b) Stereotypes. 24
    (c) Equality of respect. 25

B. LIMITATIONS OF GENETIC POTENTIAL. 30

C. DETERMINISM. 32
1. The genome and determinism. 32
2. The threat to our attitudes. 33
3. The response. 34
    (a) The rejection of fatalism. 35
    (b) Models of free action. 37
4. What would the world of hard determinism be like? 40

D. THE LIKELY IMPACT OF THE KNOWLEDGE GIVEN BY MAPPING AND SEQUENCING. 42

PART THREE: SOME THOUGHTS ON THE ROLE OF GOVERNMENT. 45

A. BASIC RESEARCH ON MAPPING AND SEQUENCING. 45

B. APPLICATIONS: GENETIC ENGINEERING. 45
1. The first phase: a moratorium. 48
2. A possible second phase: regulating the supermarket. 49

MAPPING THE HUMAN GENOME.

This report discusses ethical issues arising out of the mapping and sequencing of the human genome. The issues are of two kinds. Some arise from the possibile applications to which the knowledge could be put. Other issues arise from the knowledge itself, independent of its particular applications.

The report deals with these issues separately. Part One is concerned with what has most caught public attention: the sometimes startling "eugenic" applications to which the knowledge might be put. Part Two looks at the issues raised by the knowledge itself, independent of its applications: at the effects of the new genetic knowledge on our attitudes to ourselves and to other people.

Mapping the genome will tell us where genes are located. Sequencing will give us the order of the bases in the DNA. The two together will give us the fine structure of human genetics. But, even if the whole human genome were mapped and sequenced, other knowledge would be needed before its human importance emerged. For full understanding of the effects of genes on behaviour, we need to understand the way DNA produces chemicals and then the way those chemicals affect what people do. Little about behavior can be effortlessly read off from mapping and sequencing alone.

But, when we have the results of mapping and sequencing combined with the other parts of the causal story, we will see how genes influence physical and mental disorders, physique, intelligence, personality, etc. And this will probably give us powers of intervention and prediction.

When people talk of mapping "the" human genome, it is sometimes asked, "whose genome?". For part of the point of the study is to see the way variations in genes account for differences between people. Some of the issues of this report specifically relate to these variations: there are questions about the impact of discovering that in genetic potential we are more equal or more unequal than we suppose. But other problems considered do not concern genetic differences, but are about the impact of discovering the extent to which genes do or do not limit the options of people in general.

PART ONE: APPLICATIONS OF THE KNOWLEDGE: POSITIVE GENETIC ENGINEERING.

"Negative" genetic engineering, or gene therapy, is at present not available for safe use. When it has been tried in mice, alarming side effects have been found. But one dav we may overcome this problem, perhaps as a result of genetic mapping being complete and fully interpreted. If so, we will be able to cure disorders of genetic origin by deleting and replacing genes.

If unwanted side effects can be avoided, there is a strong case for using negative genetic engineering. Consider the options we now have when tests show an embryo or fetus to have a serious genetic handicap. We can either abort it or we can let the child be born with the handicap. Supporters of abortion in such cases would surely admit that cure would be better. And opponents of abortion would surely prefer the child to be born without the handicap. Gene therapy seems preferable to either of our present options.

"Positive" genetic engineering would go beyond gene therapy to bring about changes in people with no genetic disorder to cure. There is a boundary problem. Perhaps there is no sharp difference between altering the gene producing growth hormone in order that a person should be in the normal range rather than abnormally short, and bringing someone who would be at the short end of the normal range nearer to the middle. But, as with day and night, there is a difference even if the precise boundary is sometimes blurred.

Many people are appalled at the thought of eugenic policies designed to "improve" those who are to be born. Considering genetic engineering is a good way of getting at the deepest objections to this. For other eugenic policies raise additional problems. Attempts to influence how many children different people have, or who people will have children with, are objectionable interference in people's liberty to live their private lives as they please. Selective abortion raises the issue of the morality of abortion in general. But positive genetic engineering need not involve external interference with the parents (who could themselves choose the genes of their child) and does not raise the abortion issue. This case should bring out the central objections.

The revulsion against eugenics is often rather inarticulate. There is a danger that, if positive genetic engineering becomes possible, this inarticulacy will mean the resistance is overcome by arguments appealing to precise benefits the new techniques could bring. It seems worth sorting out the different possible benefits and the different objections before the problem is with us.

1. SOME POSSIBLE BENEFITS.

Some genetic alterations could be seen as beneficial to the person whose genes they would be. Parents might think their children would benefit from having genes making them more musical, stronger or more intelligent.

There could be benefits to society as a whole. For instance, if there were a genetic pre-disposition to violent crime, (other than crimes involving a medical disorder) positive genetic engineering might hold out the prospect of a reduction in the number of such crimes.

There is another possible case for positive engineering, which is at once more speculative and more far-reaching. If we look at the evolutionary success of the human species, it is hard not to be struck by the way our genes are still those adapted to survival in the warring conditions of the stone age. With modern weapons, willingness to fight is, to put it mildly, less obviously conducive to the survival of our species than it once was.

Many people have hoped that by political changes, such as schemes for world government or for disarmament, we may make war a thing of the past. But more pessimistic thinkers have wondered if the roots of war may be too deep in our nature for such political schemes to stand much chance of success. The relevant parts of our nature include aggression, tribalism, the ability to see members of other groups in terms of crude and hostile stereotypes, obedience, and the willingness to believe rather than doubt. If some of these characteristics are both a major contributory cause of war, and are to a large degree genetically based, it may start to seem that the further survival of our species may require changing our genetic nature.

2. RISKS AND MISTAKES.

To set against the possible gains (which tend to be more speculative in proportion to their size) there are some not very speculative risks. In the first place, there might be unpredicted genetic linkage between different characteristics. In aiming to raise the genetic limits of intelligence, we might find that we had succeeded, but that in doing so we had also brought about unexpected other changes. We might have produced people who were both highly intelligent and very aggressive and anti-social. We are used to the idea that new technologies can be expected to involve mistakes, but here the "mistakes" are individual people themselves.

There is also the point that disasters of this kind might well be irreversible. The genetic changes might be passed on to the descendents of those engineered. A society guided by humane values would not "eliminate" mistakes by killing the people it had produced in this way. And, if we attempted to limit the damage by persuading them not to have children, we might find that they rejected our suggestion in characteristically aggressive fashion.

Part of the case against positive genetic engineering is the case against risking disasters that are irreversible. In this way the objection is like that sometimes made to nuclear power, where it is objected that the nuclear waste problem is one which we are leaving to plague generations thousands of years from now. Optimists reply that they will have developed the technology for dealing with it, while others reply that only an extreme emergency could justify our leaving such a massive problem to future generations, who we only hope will have the means to cope.

In the case of positive genetic engineering, we are contemplating improvements rather than reacting to an emergency. And, as we have seen, the putting right of mistakes raises dilemmas which cannot be overcome merely by technological advance, but which require persuading people that their genes should not be passed on to their descendents. In these ways, the case against risking disaster through positive genetic engineering is stronger than that against leaving nuclear waste.

It could be said that genetic mapping, when both complete and fully interpreted, would give us the knowledge of the relevant genetic links and so enable us to avoid these disasters. But there is still the question of whether we might be wrong in thinking that we had a full interpretation of the significance of our map. And there is also the difficulty of excluding human error in applying even a technology based on a very full theoretical understanding.

Another problem would be the danger of reducing the evolutionary resilience of the human gene pool. This could happen in two ways. We might delete particular genes which have entered the genome because they make a contribution to suvival which we do not appreciate, perhaps involving resistance to a disease. Or we might narrow the variety of the gene pool by our choices. One evolutionary strategy for the survival of a species is to have a very varied gene pool, so that a new disease which destroys many of its members does not destroy them all. We might risk removing this safety net for species survival.

We could lower the dangers by limiting the numbers or the kinds of genetic changes people could make. But, to the extent that we do this we also reduce the benefits which can be expected from positive engineering.

3. WHO WOULD DECIDE?

A major problem this technology will bring is who is to have the power to use it. Many discussions of it are in terms of some group of scientists deliberating on the acceptable applications. This can be alarming because of the possible limitations of human variety which might result. The thought of our genetic future being determined by the limits of imagination of the members of any one profession is a rather narrowing vision. The committee could of course be not just a scientific one, but could be more broadly based, including others thought to be knowledgeable and wise. But even if the committee included mothers, doctors, lawyers, philosophers, a rabbi, a minister and a priest, we might wonder about handing over to them the genetic future of the human race.

The drawbacks of the "central committee" model have led to other proposals. One of these is what Robert Nozick (in Anarchy, State and Utopia) has called the "genetic supermarket". Here the choices would be left to parents. They would simply be told of the available possibilities, and, if they wished, they would within those limits choose the genetic characteristics of their children.

The genetic supermarket has several advantages over the central committee. It seems more likely to preserve human variety. And, to some of us at least, it seems that parents are much more to be trusted to take decisions for their children than is a committee to take decisions for other people's children. The seriousness and sense of responsibility they bring to the decision may usually be much greater.

It is worth noticing that some of the possible benefits of genetic engineering would be less likely to be available on the genetic supermarket version. If we are thinking in terms of changing human nature in ways to eliminate war, this seems to require changes which are virtually universal.

If only some parents changed the psychology of their children in the appropriate ways, the problem might not be solved. The human race might just divide more sharply into those who made the wars and those who were their victims. (A parellel problem arises between countries. Even if one country thought of making the changes universal in its own next generation, they might be put off by the knowledge that some country traditionally a rival of theirs had decided against such change.)

And, apart from foregoing some possible benefits, the genetic supermarket has some internal problems of its own. There is the question of how far parental freedom of choice should be restricted in the interests of the children. Most societies accept the case for giving children protection against some harmful parental decisions. For instance, many countries do not allow parents to refuse schooling for their children, or to decide against their receiving various kinds of health care. Decisions would need to be taken about whether some kinds of genes, like some drugs, should be excluded from the supermarket shelf.

Exclusions from the supermarket might not just be on grounds of harming the child whose genes are in question. There would be questions about the interests of society as a whole.

Here the issues parallel those raised by technology with a less science fiction flavor. Either through abortion or through more subtle techniques now developed, it is possible for parents to choose the sex of their child. There is a danger of a massive sex imbalance in the next generation if these techniques are widely taken up. No particular child is harmed by being one sex rather than the other. (There are two reasons: the child of the opposite sex who might have been born would not have been that child, and -apart from possible prejudice -it is not worse to be one sex than to be the other.) But, while the child whose sex is chosen may not lose by that, the imbalance may make people as a whole worse off. For one thing, roughly even numbers make people's sex lives easier.

It is easy to see that, if some genotypes became fashionable, the genetic supermarket could lead to social imbalances of a similarly worrying kind.

There is also a danger of a kind of genetic competition between parents which would lead to a state of affairs with more losers than winners. Suppose some parents chose to make their children more thrusting and competitive. Others might feel that, unless they did the same, their children might lose in competition with the thrusting ones, and so feel compelled to follow suit. We might end up with a society which most of us would not have chosen in preference to the present one. (As Fred Hirsch said about some aspects of economic growth, "where everyone stands on tiptoe, no-one sees better". Social Limits to Growth, London, 1977.)

If positive genetic engineering is to come, the genetic supermarket is a much more attractive starting point than the central committee. But it is only a starting point, from which at best a satisfactory decision procedure will emerge by thinking through the restrictions needed to protect the children and the interests of the wider society. In a democracy these are political issues. Elections in the twenty-first century may be about more exciting issues than economics.

4. JUSTICE AND EQUALITY.

If the genetic supermarket is literally that, and is run commercially, "better" genes are likely to go to those whose parents can pay more for them. If this happens, children who grow up in relative poverty will suffer a double disadvantage: they will have the present handicaps of being poor, together with the more fundamental one of knowing that their abilities have narrower genetic limits than those of rich children. This would involve as strong a denial of equality of opportunity as can be imagined. It would be a move towards a society of genetically separated castes.

A similar problem exists on a world scale. With most technology, the advanced countries benefit from it first, and third world countries lag behind. If this happened with genetic engineering, the trailing behind would be compounded by the people of those countries themselves being more limited than those in the advanced countries. For a technology gap between nations to be supplemented by a gene gap does not look like an advance towards global justice. Nor does it look like a recipe for a stable and peaceful world.

When considering equality of opportunity, whether within a nation or between nations, two confusions are to be avoided. First, genetic inequalities need not be either more extreme or morally worse than environmental ones. The bad effects of growing up in severe poverty or in a very unhappy family may be as great and as ineradicable as those resulting from genetic disadvantage.

The second point is that for some people to have their genes chosen while others do not, need not in itself create any greater inequality of opportunity than would otherwise exist. The natural genetic lottery seems to involve a good deal of inequality of potential. And an inequality does not become greater because some of the unequal characteristics have been created rather than have come about naturally. Degree of inequality is independent of causal origin. What leads to an overall worsening of inequality of opportunity is when genetic inequality starts to be linked with social and economic inequality (or when any such link that already exists is increased by the use of money in the genetic supermarket).

There is also a problem about the impact of genetic engineering on equality of respect. At present people have a huge range of different characteristics. Sometimes people envy those others have. But, to a great extent, differences are not subject to a uniform rank ordering. Being good at tennis is not in some competition with being a good architect. But where it became common to engineer some characteristics in preference to others, it would be hard to avoid some spill-over to the evaluation of people already born with those characteristics.

If genetic engineering presented us with a choice between some central abilities of a good architect and those of a good tennis player, and people showed a massive preference for their children having those of the tennis player, it might be hard for architects not to experience some loss of self esteem. And they might also suffer from some loss of public respect, as people saw that their skills were not widely ranked high.

The question of the nature of equality of respect is discussed in part two of this report, but it is at least worth mentioning here this possible erosion of one version of it as a result of genetic engineering. This erosion may only accentuate something which already exists. There is already a degree of socially understood ranking of the skills associated with different jobs, which already leads to some inequality of esteem among people. Genetic engineering would add to this by for the first time putting genetic characteristics directly in competition with each other. And it might extend the scope of the ranking to characteristics barely touched by it now.

5. AUTONOMY AND RESPECT FOR PERSONS.

One objection to positive genetic engineering is often put in the form of an objection to people "playing God". Sometimes this is based on an explicitly religious view: it may be held that there is a divine plan, which genetic engineering would blasphemously interfere with. (People who take this view have the problem of whether any interference with nature frustrates the divine plan. If so, they have to oppose medicine as well. If not, they have to explain how they know which kinds of interference are allowed and which ruled out.)

But the "playing God" objection is often not a distinctively religious one. The form of words can express a resistance more widely held than among believers in a divine plan. It may in part refer to the dangers already mentioned, of disasters resulting from our fallible understanding. (We lack the omniscience which some may feel would be a precondition of positive genetic engineering.) But the "playing God" objection is often more than this. It is often a challenge as to whether anyone has the right to decide what someone else's characteristics shall be.

In this form the objection is one about respect for persons. If I take it upon myself to engineer your genetic characteristics, I am treating you as my creature, and this shows a lack of the respect due to you as a person.

The idea of respect for persons does seem to capture some of the force of the intuitive rejection of positive genetic engineering. But it is an elusive idea, which in other contexts has often seemed both apt and at the same time hard to pin down. Some aspects of respect for persons are discussed in part two of this report. But one aspect seems particularly relevant to mention here: respect for persons seems to require, among other things, respect for their autonomy. Slavery is a denial of respect for persons, even if the slaves are well treated and are happier than they would have been if free, because their autonomy has been removed.

Does genetic engineering of someone's characteristics violate that person's autonomy? Intuitively, we may feel that it does. But there are some grounds for doubt. The idea of violating my autonomy carries with it the implication that my desires or my choices have been over-ridden and that I am less free. But genetic engineering never involves the first, and need not involve the second.

The person whose genes are being engineered is not at the stage of having desires or of making choices. So there can be no question of autonomy being violated by desires or choices being over-ridden. Such a claim about autonomy is as implausible as it would be to object to abortion on the grounds that the fetus wants to go on living.

And, although it is natural to object to genetic engineering on grounds of freedom, this too need not apply. For it is not clear that someone whose genes have been engineered has any less freedom than someone whose genes have not been interfered with. For my freedom is partly a matter of what things I am able to do. And engineered genes may impose limits no narrower than natural genes. It is more likely that genetic engineering will have increased my capacities. A genetic engineer would have to be incompetent or malevolent to have chosen genes that systematically made me weaker, more stupid and less skilled than I would have otherwise been.

Part of the case for positive genetic engineering is that it can increase people's freedom by extending their abilities and so enlarging the options open to them. And it does not involve the denial of autonomy which consists in over-riding someone's own desires or choices.

So one defence of positive genetic engineering against the charge that it reduces autonomy is to show that it may increase the choices open to people. But this defense may lay it open to a quite different charge. Is it clear that a wider range of choices is a benefit rather than a burden?

There is for instance Kierkegaard's view that "increasing options may increase despair". Perhaps despair as a reaction to greater numbers of options is rather exaggerated. This response might not be found often except in Kierkegaard and his twentieth century existentialist followers, who have perhaps excessively romanticised the agonies of choice. But there can be drawbacks to having greater options to choose between for one's children, not least because of the burden of responsibility, and the worry about making the wrong choices for them. Positive genetic engineering could give a new twist to teenage criticism of parents: "why didn't you give me better football skills and blonde hair, instead of a higher IQ and this terrible black hair?"

Despite the extra responsibility and the problems which come with greater choice, it is hard to see restricted choice as preferable. We normally think it is better if parents have a choice of schools to send their children to rather than just one, despite the extra responsibility this brings. And a case would have to be made out for seeing genetic choice as different in principle from choices such as this. Without such arguments, it seems reasonable to count wider options as a gain rather than as a loss.

Although positive genetic engineering is more likely to increase people's choices than to diminish them, there is a persisting feeling that some objection based on autonomy does apply to positive genetic engineering. The reason for this tension in our thinking seems to be that "autonomy" has two separate components, usually interwoven, but which pull apart in the case of genetic engineering. One component is being free to take and implement decisions about my own life. The other is a matter of my life not being dominated by other people.

These strands pull apart here because, when compared to having natural genes, having genetically engineered ones does not leave me less free to take or implement decisions about my life. I do not choose my genes on either system, and having engineered ones may actually extend this control. But, on the other hand, my life, and what I am, is dominated by the person who chose my genetic characteristics.

In one sense, genetic engineering preserves and extends my autonomy, while in another sense it reduces it by placing me under the dominance of another person. Different people may give different weight to these two different aspects of autonomy. But there is a case for saying that the second aspect is relatively unimportant. And this is that it makes little difference to my situation whether my genetic nature is formed by the genetic lottery or by someone else's choice. In either case it is outside my control: something I must take as given.

6. IMPLICATIONS.

Many would want to exclude positive genetic engineering in principle. The upshot of this discussion is more complex. There are enormous dangers in positive genetic engineering. There is the risk of adding "mistakes" to the gene pool which it may not be possible to erase. These "mistakes" will be people with messed up genes. There is the danger of commercial exploitation leading to gene-linked inequality of opportunity. And there are many unsolved problems of how the decisions about the use of the technology are to be taken.

All these considerations add up to a strong negative case. But are they enough to support the claim that it should be ruled out in principle for ever? I am skeptical about this. One reason for skepticism is that it is hard at this stage to be clear what benefits positive genetic engineering may offer to offset the drawbacks and problems. It is also possible that we will start to see ways round the problems, which will as a result become less daunting objections. And, perhaps, the possible links between war and our present genetic nature should not be forgotten. You have to avoid thinking too hard about history, including that of the twentieth century, to be very confident about ruling out possible ways of changing what we are like.

But, this having been said, any benefits are highly speculative, and the problems are very real. Those of us who are not prepared to exclude positive genetic engineering absolutely and in principle are hardly enthusiastic about rushing in to use it as soon as it becomes available. For better or worse, starting to change our genetic nature could be one of the turning points of human history. And we have hardly begun to think through the problems it raises, particularly about who should take the decisions and what restrictions should be imposed on the choices available.

This adds up to a case for at least giving ourselves a breathing space in which to think about what is and is not acceptable. If positive genetic engineering starts even to appear on the horizon, we should impose a moratorium, legally backed if necessary, to ensure that it is not used until we have had time to decide if it should be used. In that way, we may for once ensure from the beginning that a new technology is harnessed to our values instead of running away out of control.

PART TWO: THE KNOWLEDGE ITSELF: ITS IMPACT ON ATTITUDES TO PEOPLE.

Mapping and sequencing the genome raise issues independent of the applications of the understanding we would gain. The possession of the knowledge itself, quite apart from what we did with it, would in the long term have a deep social impact.

A great increase in our understanding of human genes and of how they shape our characteristics could influence our attitudes to ourselves and to others. It could do so by giving us "diagnostic" knowledge of individuals, whether of medical abnormalities or just of iddividual differences between "normal" people. It would be likely to highlight the degree to which the genetic potential of different people is equal or unequal. A second kind of knowledge would not be of individual variation, but of the impact of genetic factors on people in general. This might show us that we have fewer options than we suppose, and perhaps encourage a determinist view of human choices.

A. EQUALITY AND INEQUALITY OF GENETIC POTENTIAL.

Most of us have some vague idea that, for many of our characteristics, our genes set limits to what we can be, without totally determining it. I may be fairly sure that, with my genes, nothing could have made me a musical prodigy or a champion athlete. But, if I had had a more musical upbringing, or had taken more exercise, I could have been a bit less bad in these fields than I now am. Most of us have beliefs of a rough kind about our genetic potential in such spheres: about the limits of possible improvement through effort or environmental change. And we have similar beliefs about other people.

These intuitive beliefs are vague and often poorly based. Inevitably, it is only with a few skills or characteristics that we have pushed against the limits of our potential. We may be often be wrong about what our own limits are. And our evidence about others is usually even hazier. So when science enables us to trace people's actual limits, our intuitive picture may often turn out to be wrong. The correction could be in either direction: we may have greater or less potential than we thought. (The correction could be in different directions for different characteristics.) And the same holds for differences between different people. Differences of potential may be greater or less than we suppose. (Again, the correction may go different ways for different characteristics.)

1. THE IMPACT OF DISCOVERING THAT WE ARE GENETICALLY MORE EQUAL.

Take some central skills which contribute to educational success, for instance the cluster of skills measured by I.Q. tests. The working assumption of many people is that, to a degree not yet precisely measured, both genes and environment contribute to individual variation in these skills. Suppose it turns out that the variation in genetic potential is much less than has been thought: that the differences are overwhelmingly to be exlained in terms of environmental factors.

When this became widely understood, one likely result would be to build up much greater pressure for equality of opportunity. Where differences of educational attainment are seen as partly genetic, this weakens the sense that inequalities of outcome are unfair. It can be thought that improved opportunities might not have led to a very different result. But this thought is removed when we know that people's genetic potential is close to equal.

It should be said that, just because a difference is not genetically determined, it does not follow that it is the result of factors which can be changed by social policy, such as better schools. Factors such as order of birth in the family, or coming from a family where people discuss things a lot, may cause differences as resistant to public policies as genetic ones. But it seems likely that many environmental factors will not be so resistant. The removal of the genetic factor will make removable environmental disadvantage seem more obviously unfair in its effects and more obviously something to be removed.

If that is true at the level of differences between individuals, it is also true of differences between groups. Where differences of ability are correlated with sex or race, the thought is now open to people that this may be the result of differences of genetic potential. Demonstration that this is false would again be likely to expose those environmental disadvantages that are alterable (including those related to sexism or racism) to greater pressure.

2. THE IMPACT OF DISCOVERING THAT WE ARE GENETICALLY MORE UNEQUAL.

(a) ALLOCATION OF RESOURCES. Take again the cluster of skills measured by I.Q. tests. Suppose we found that environmental influences had negligible impact on individual differences, because genes had a very rigid determining influence on them.

This knowledge could pull two ways with regard to the distribution of educational resources, where we also knew which people higher and lower potential. It could lead to a compensatory approach, giving greater priority to ensuring that those with lower potential reached their ceiling than to doing the same for those with higher potential. This would be an effort to reduce the gap between the achievements of people below that suggested by the differences of genetic potential. Or our knowledge of these differences could pull the other way, with a smaller share of resources going to people who seemed a bad bet.

(b) STEREOTYPES. Where great differences of genetic potential correlated with membership of socially visible groups sush as race or sex, there would be a real danger of the knowledge creating or reinforcing stereotypes. Although the difference might only be an average one, with many individuals in the "lower" group having greater potential than many in the "higher" group, there could well be a tendency to write off all members of the "lower" group when thinking of the skill in question. And the effects of this would be worse to the extent that a group was "lower" across a whole range of characteristics.

(c) EQUALITY OF RESPECT. If some characteristics come high up in many people's ranking of those that are most desirable, this may spill over to a ranking of people according to their differences of potential for them. Perhaps your genes give you a much greater possibility of being articulate, attractive, good at mathematics, science and sports, compared to the narrower limits my genes have imposed on me. My genetic potential may surpass yours to the same extent for other characteristics: I may have an amazing memory for numbers, remarkable speed and accuracy in felling trees, and an ability to recognize as an individual each of hundreds of sheep. But if society values the characteristics where you score much more than those where I do well, it is possible that the impression may get about that you are in genetic potential a more high-powered person than I am.

There is likely to be some ranking of people according to their potential score for the "best" characteristics. For those of us who are unenthusiastic about people being ranked on a single scale, it is fortunate that there is a plurality of values: we do not all agree on the ranking of characteristics. But some characteristics are sufficiently widely thought better or more important than others for there to be at least a tendency towards the single scale.

With this tendency, there co-exists a widespread ethical belief in equality of respect. Many of us feel that there is some deep equality between people, and this feeling issues in a belief that, in some respects, equality of treatment is required. (This is part of the explanation of our appalled response to the Nazi view that some people are sub-human, or of the appalled response we now have to the way many slave-owners thought of their slaves.)

To what extent is equality of respect threatened by a clear view of inequalities of genetic potential? There is a benevolent view that no threat exists, because at some deep level we all do have equal potential.

This view is found among many of the best schoolteachers: one child may have more mathematical potential, another may be more musical, another may have the possibility of being a terrific footballer, another may be a wonderfully imaginative storyteller. The point is that every child has something, and when all these things are seen together, you cannot say that any child lacks potential compared to any other. This is obviously a marvellous view for a teacher to take, and it must lead to a lot of hidden potential being sought out and developed. It must often lead to children developing a lot of mutual respect.

This benevolent view has such good results that I hate to suggest that it is probably in part an illusion. But it is.

The truth in it is that most people have many good things in them, and that what they have to contribute is varied in ways that defeat a single ranking. The illusion is that the qualities people have must somehow even out so that we are all equally endowed.

Consider abilities within a given field: sports. There could be a world in which everyone had equal sporting ability. Some are good at football, some at baseball, some at running, some ay gymnastics: but it all evens out to equal overall ability. Our world is not like this. Some are close to being good all round, and some are close to being bad all round. When we return to the broader perspective, and look at all kinds of potentiality, is it not just benevolent optimism which makes us assume that there things are not like in sports, but that there is a kind of pre-established harmony such that we are all equal in potential?

If this "benevolent illusion" is an illusion, there seems to be a dilemma for equality of respect. It seems to rest on some belief in people being in some ways equal, and in virtue of this "qualifying" for equal respect. The dilemma is this. Either the relevant ways in which people are equal are their empirical characteristics (their talents, their warmth, their virtues, their intelligence, their charm), or they are equal in some more metaphysical sense (they have a soul, or they are "equally human"). It is a dilemma because, on the one hand, the empirical equality seems to be an illusion, and the metaphysical equality seems so nebulous and elusive that it is not clear what a belief in it comes to.

If both the empirical and the metaphysical versions of the belief that people are equal are unsatisfactory, this suggests an alternative: perhaps it is a mistake to think that equality of respect needs to be "grounded" in a belief of this sort at all.

It is worth distinguishing two senses of "respect". There is respect we have for people, in virtue of their characteristics or what they do: the respect we may have in greater measure for some of our friends than for others, or in greater measure for Mother Theresa than for Hermann Goering. (This kind of respect is sometimes called "esteem".) And there is respect we show people, which is a matter of how we treat them, and of a certain attitude we adopt towards them.

Showing respect for people has two aspects. It requires that we respect their rights, for instance that we do not assault them, steal their property, tap their telephone, or prevent them from expressing their religious or political beliefs. (There is a problem of how to decide what people's rights are, but that does not affect the claim that showing people respect requires not violating their rights, whatever they are.)

The other aspect of showing respect is harder to pin down, but hardly less important. It is a matter of attitude. It is a matter of not treating people with contempt or condescension. It is absent, to different degrees, in a caste system, or in a snob, or in someone who never listens but always interrupts. It is what makes our attempts to teach children to be courteous and considerate not just a matter of indoctrinating them in a set of arbitrary rules.

There is a tendency to muddle having respect (in the sense of esteem) for people with showing people respect. And where they are seen to be separate, there is a tendency to believe that the latter presupposes the former. This makes it seem that equality of respect is based on the benevolent illusion that there must be some way in which all people ultimately will turn out to have equal potential or to be of equal merit.

But, of course, we can respect the (equal) rights of someone whose merits seem to us to be less than those of some others. And having this poor view of them need not exclude treating them in a considerate way. Advances in understanding of genetic potential may help to shatter the benevolent illusion. For those of us who care about equality of respect, it is worth making clear now that what we care about does not depend on that illusion.

It could once be held to be a self-evident truth that all men are created equal. It is certainly not now self-evident, and may be about to be shown to be false. But another American belief, in equal concern and respect, is not tied to the self-evident truth, and can and should survive its demise.

B. LIMITATIONS OF GENETIC POTENTIAL.

New genetic knowledge may affect our attitudes in ways not related to how far we do or do not differ from each other. One way in which it could do so is by showing us that genes limit the options open to all of us far more than we now suppose.

There are some striking anecdotal accounts of identical twins separated early in life, who meet up as adults. Our intuitive thoughts about the influence of genes might lead us to expect them to look very like each other and to have some very general similarities: to be very close to each other in I.Q., etc. But in a number of cases of twins separated at birth, what was striking when they met for the first time as adults, was not so much a general similarity of temperament, but the way small quirks would turn out to be the same. Sometimes they would have given their children the same names, or both would have built a white seat round a tree in the garden, or both would obsessively count the wheels on passing trucks, or both would be afraid of the sea and would go into it carefully backwards, or both would have a habit of pushing up their noses, which both would call "squidging". (Peter Watson: _Twins_, London, 1981.)

These cases, provided that they are not merely coincidental, suggest that genetic influences on behaviour can be surprisingly localised and specific.

It seems possible that understanding the fine structure of genetic influence on behavior will replace anecdote with harder evidence, and will show this kind of detailed genetic control over parts of the behavior of all of us. If things turn out this way, the knowledge is bound to revise our picture of the range of options we have. Our intuitive view now does not, for instance, include the thought that our genes may make our choice of children's names virtually inevitable.

One present idea we have, which is of some importance to us, is the belief that, at least in part, we can create the sort of person we are.

There seem to be three interlocking processes. There are conscious projects of self-creation: we aim to make ourselves more generous, better organized, less overweight, or whatever it may be. Then there are decisions not motivated by any picture of what we want to be like, but which yet make a difference: decisions about who to live with or marry, where to live or what job to take. And then there is the slow unfolding of the genetic program. We do not choose the emotional turbulence of adolescence, or the slow decline of middle age and beyond. These seem as unavoidable as the physical changes of the ageing process. Greater knowledge about genetic control may make this third process have a much wider scope than we now suppose.

This might have the effect of making us give up some attempts to change either ourselves or other people, which would be seen to be pointless. The effect of such a change of outlook would be at first disconcerting. But it would not be disastrous, unless it led to a general fatalism, which made _any_ attempt to change anything about ourselves or anyone else seem futile. Although the new genetic knowledge may contribute to a determinist view of people, it can be argued that this fatalism is not a necessary consequence.

C. DETERMINISM.

At present, we are enough influenced by the scientific belief that causal mechanisms are to be found to half believe in a determinist picture of people. But we also half believe that many choices are free in some sense that eludes causal predictability. But, as the causal mechanisms, including the genetic ones, are filled in, there may seem less and less room for this "contra-causal freedom". We are likely increasingly to be pushed towards a determinist picture.

1. THE GENOME AND DETERMINISM.

Mapping and sequencing the genome will not alone impose a determinist picture on us. Seeing where genes are located, or knowing the order of bases in the DNA, will not alone make behavior predictable. Knowing the ground plan of a factory and the internal structure of the machines in it gives only clues about the numbers and kinds of cars it will produce.

But mapping and sequencing together with tracing the pathways between genes and behavior will start to fill in the determinist picture. We are now starting to work out these pathways. Take for instance the pattern of behavior classified by psychiatrists as "sensation seeking", which involves a disposition towards gambling and alcoholism. We now know that this is correlated with low levels of activity of platelet monoamine oxidase. (Buchsbaum et al.: The Biochemical High-Risk Paradigm: Behavioral and Familial Correlates of Low Platelet Monoamine Oxidase Activity. Science, 1976; Von Knorring et al.: Personality Traits Related to Monoamine Oxidase Activity in Platelets. Psychiatry Research, 1984.) And those levels of activity have in turn been shown by twin studies to be largely under genetic control. (Nies et al.: Genetic Control of Platelet and Plasma Monoamine Oxidase Activity, Archives of General Psychiatry, 1973; Revely et al.: Genetics of Platelet MAO Activity in Discordant Schizophrenic and Normal Twins. British Journal of Psychiatry, 1983.)

In such a determinist picture, our actions can all be explained in terms of (no doubt very complicated) causal mechanisms. If this model is right, it seems that what we do, just as much as what billiard balls do, is the product of a set of causal laws operating in particular circumstances.

2. THE THREAT TO OUR ATTITUDES.

The determinist picture seems disturbing. It suggests that a God-like scientist, with complete knowledge of all the relevant causal laws, and of the circumstances in which they operated, could successfully predict any human action. If anyone did anything other than the predicted action, this would falsify one or more of the causal laws. But, the argument goes, determinism requires that some set of relevant causal laws is true, and this is incompatible with their being falsified. So determinism is incompatible with anyone not doing the predicted action. Ruling out all courses of action but one, it seems incompatible with genuine choice.

In two different ways determinism is at least an apparent threat to our attitudes. First, the elimination of genuine choice would leave no room for choosing what sort of person to be, and so our belief that we can (partly) create ourselves is threatened.

The second threat is more oblique. Undermining choice may also undermine many of our emotional reactions to people. The determinist picture may not leave room for justifiable resentment of what people do, or for justifiable feelings of blame or guilt.

3. THE RESPONSE.

If determinism is true, we are left with the choice between what William James once called "hard" and "soft" determinism. Hard determinism is the disturbing line of thought we are considering: the view that determinism rules out genuine choice, and so also rules out the emotional responses linked with holding people responsible for what they do. Soft determinism asserts that free choice and responsibility are compatible with determinism. (In some versions, they even require it.) Soft determinism (sometimes known as "compatibilism") assumes, just as much as hard determinism (sometimes known as "incompatibilism"), that in principle actions are completely predictable. It is not "soft" because of any fuzziness in the causal laws.

The issue is whether the soft determinist can resist the hard determinist argument against freedom and the reactive attitudes. Soft determinists have developed two main alternative strategies for doing this. The first strategy points out that determinism is not the same as fatalism: that, even in a determinist world, what we do influences the future. The second challenges the claim that determinism eliminates genuine choice. The aim is to work out a model of free action that is compatible with determinism.

(a.) THE REJECTION OF FATALISM.

A fatalist thinks that there is no point in my doing anything, as the future will be as it will be, regardless of what I do. This is a silly view, whether or not determinism is true. It has the consequence that I may as well drive when drunk, since, if there is going to be an accident, nothing I decide can make the slightest difference.

It may be agreed that fatalism is intuitively highly implausible, but it may still be wondered how a believer in determinism can escape from it. If, as determinism suggests, everything is in principle predictable, does not this mean that everything, including my present and future actions, and their consequences has already in some sense been decided? And, if so, is not it an illusion to suppose that my decisions and actions can make any difference to that already predictable future?

It is true that in a determinist world, future events are in principle predictable. But this is because they are the product of the causal factors which shape them. And these causal factors include the decisions and actions of people. The road accident is predictable, but on the basis of prior causes, such as my decision to drive when drunk. My decision is important, even though it is itself in principle predictable. If it goes one way, there will be an accident, and if it goes the other way there will not be one.

In a determinist world, outcomes are the result of earlier causes, and there is no reason why my decisions and actions should be excluded from those causes. So I am justified in deliberating and acting "as though" the actions I choose to perform make a difference to the world. Because they do make a difference. And there is something paradoxical and absurd about using the truth of determinism as a reason for deciding never to take decisions on the basis of reasons.

(b.) MODELS OF FREE ACTION.

The hard determinist argument often relies on the words "can" and "could". It says that, if what I do is predictable on the basis of causal laws, then I cannot avoid doing it. If what I did yesterday was predictable in the same way, I could not have done otherwise. Determinism seems to entail that I cannot do anything except the things I do. Soft determinists, challenging the argument, have objected to this step. The objection appeals to how we ordinarily think about abilities: what we mean when we say that we can or cannot do things. The objection is that there is a perfectly good sense of "can" in which I can do many things I do not do. And, similarly, there is a perfectly good sense of "could" in which I could have done many things I did not do.

A classic statement of this objection was by G.E. Moore. He said, "I could have walked a mile in twenty minutes this morning, but I certainly could not have run two miles in five minutes", and "It is true, as a rule, that cats can climb trees, whereas dogs can't". (Ethics, London, 1912, chapter six.)) (I like "as a rule". No-one is going to catch out Moore by citing some cat with amputated legs.)

This case for soft determinism starts with an account of what it is to be able to do something. I am able to walk a mile in twenty minutes, in that nothing would frustrate any decision to do so. And I have this ability even if I do not choose to use it. None of this is upset by the truth of determinism. And the presence of such abilities is all that is presupposed when we hold people responsible for what they do.

But this is at best only a first stage in constructing an adequate soft determinist account of free action. Suppose there is a sense of "can" or "could" which is to be explained along the lines Moore suggests. Perhaps there is also another sense in which "he could have done otherwise" is excluded by determinism. (It is the sense in which, given the laws of physics and the forces bearing on them, billiard balls cannot move anywhere except where they do.) Do we know that this other sense can be ignored when we are thinking about someone's freedom to act?

There is another doubt about this version of soft determinism. It makes a sharp separation between desires and abilities. This may not be a problem when we are concerned with physical abilities. But, with psychological abilities, the separation seems less clear. Consider possible causes of "diminished responsibility", such as alcoholism or compulsive desires to steal. In these cases, there may be other blocks to freedom, as well as those which come between choice and action. Perhaps choices themselves can be unfree. And, even apart from these special cases, it is reasonable to ask, "I can do what I choose, but can I choose what I choose?" Or, as the question is sometimes put, "I can do what I want, but can I want what I want?"

In reply to this, a more sophisticated soft determinist account has been developed by Harry Frankfurt (Freedom of the Will and the Concept of a Person, Journal of Philosophy, 1971). On his view, we do not just have desires to act, but higher-order attitudes towards those desires. There are second-order desires: someone can want not to have his desire to smoke.

Frankfurt calls tne first-order desire which is actually expressed in choice the person's "will". He says that freedom of action is being able to do what you want. (It is what Moore was concerned with.) But, for Frankfurt, there is also freedom of will: the ability to have the will you want. He thinks that these two freedoms together make up the whole of freedom. He says of someone who has both: "Then he is not only free to do what he wants to do; he is also free to want what he wants to want. It seems to me that he has, in that case, all the freedom it is possible to desire or to conceive."

There is a problem for this sophisticated account of freedom of will, parallel to the problem faced by Moore's one-level account of freedom of action. At whatever level of desire we stop, can we not ask whether that desire was avoidable? There seems a danger of having to go endlessly further back, asking whether someone desires to desire to desire to act, and so on. All this has to start with some set of attitudes and desires that we just take as given. And the content of the set we start with is the product of our genes and early environment, neither of which we could choose. What we start with is just good or bad luck. And perhaps it is unfair to blame people for the results of luck in the genetic and environmental lotteries.

Hard determinism has not been defeated. It is not clear that we _have_ to abandon blame, etc., but it is not clear that we are going to be able to justify them either. In this situation, we may be wise to look at what it would be like to move over to a world in which the hard determinist view prevailed. There are reasons for thinking that the change might be less drastic than we might think.

4. WHAT WOULD THE WORLD OF HARD DETERMINISM BE LIKE?

The threat is to the reactive attitudes, such as blame. It may seem that, in a hard determinist world, these attitudes would either wither away, or linger on in a feeble form, like superstitions no-one really believes in.

But it is wrong that the only responses to people in a hard determinist world will be cooly scientific. Reactions which presuppose desert will have to be given up. But a whole range of reactive attitudes are not desert-based. Consider the aesthetic-cum-sexual responses we have to people's appearance, or to their style and charm. We have aesthetic responses of another kind to people's intellectual qualities: to their being imaginative, independent, or quick on the uptake. These responses are not desert-based. We do not think people are attractive, or quick on the uptake, because of praiseworthy efforts they have made. On the contrary, these seem very clear cases of features owing a lot to luck. Perhaps this is one reason why we think of our responses to them as aesthetic rather than moral.

These responses will save the hard determinist from a totally dry emotional life. There seems no reason to stop at that point. There can also be aesthetic responses to people's motives and character. If I am a hard determinist, I can still admire someone's generosity, so long as I accept that this is in part good luck. It will be like admiring someone's musical talent. And, when someone swindles me, I can have similarly unfavourable responses to their character. My judgement that someone is repellantly selfish in outlook will be like the judgement that someone has a repulsive smile.

These aesthetic responses could be conveyed to other people. And, just as people want to be thought physically attractive, rather than ugly, so people would care about aesthetic responses to their character. Desert-based attitudes, like blame, would have been renounced. But unfavourable aesthetic responses to character as revealed in actions might function like blame. They might be just as effective in putting pressure on people to change their behaviour.

But it is hard to see how retributive punishment could be justified when hard determinism was accepted. At present, we justify penal policy partly on grounds of its hoped impact in reducing the crime rate (through reform or deterrence) and partly on the grounds that offenders deserve to be punished. This last appeal would have to go.

Someone's bad character may provide a ground for my negative aesthetic reaction to it. But, where it is accepted that, in an important sense, that character could not have been different, its badness looks a less good ground for imposing a penalty. A penal policy based entirely on consequences looks like the only justifiable one. (Such a policy would still allow that innocent people should not be punished, but the basis for this would be a consequentialist one.)

D. THE LIKELY IMPACT OF THE KNOWLEDGE GIVEN BY MAPPING AND SEQUENCING.

It is worth emphasizing again that mapping and sequencing of the human genome will on its own change little. It is only when this knowledge is interpreted in the light of greater knowledge of the chemical pathways linking genes with behavior that we will see how genes do or do not shape and limit what we can do. But when we do have this understanding, we may be able to read off from someone's genome a great deal about his or her physical and psychological characteristics, together with a lot about his or her potential and limitations.

This fuller understanding of the impact of our genes on our lives would over time be likely to have a deep impact on the way we see ourselves and other people.

If we discover that our potential is more equal than we now think, this may lead to more pressure for equality of opportunity. If we discover that we have greater inequality, this may lead to compensatory environmental efforts, or, more depressingly, it may lead to individuals or groups being more readily written off. Knowledge of greater inequality at first sight seems to threaten the ideal of equality of respect, but if we are clear-headed we will not give up the ideal.

If we find that, in general, fewer options are available to us in our lives than we have supposed, this will be a substantial modification of our self-image, but need not lead to a passive or fatalistic approach to life. If the new genetic knowledge gives support to a generally determinist view of people, this will again have a large impact on our attitudes, both to ourselves and to other people. Our emotional responses to each other may not feel very different, but some of them will have a different basis. And we are likely to come to see retributive punishment as unjustified.

In matters of this kind, the changes brought about by new ways of seeing things are slow. There will be no sudden switch: rather a deep but gradual change of view. In this respect, the change of outlook brought about by Darwin may be a model.

From different points of view, various of the changes predicted here may be viewed with approval or disapproval. But even if one's overall response is unenthusiastic, this is not a sufficient reason for attempting to stop the research. These changes of attitude do not add up to a cogent case against genetic mapping and sequencing in themselves. To the extent that our attitudes are based on ignorance or illusion, they are themselves unjustifiable. If so, unless we give up on rationality altogether, it is better that they should be changed. And it would be self-defeating to try to shore up our present attitudes by stopping the research which might undermine them. If this were public policy, we would see that it was happening, and the attitudes being protected could hardly survive such a defense unscathed.

PART THREE: SOME THOUGHTS ON THE ROLE OF GOVERNMENT.

The scope and limits of justifiable government intervention is often controversial. Much of nineteenth and twentieth century political debate has been between different views of the boundary between the public and private spheres in the economy. We can hardly expect this boundary to be less controversial in the context of technology which may enable decisions to be taken about the kind of people there should be. In a democracy, these matters are "settled" by elections. But this should encourage, rather than exclude, attempts to think out a defensible view.

There is a case for a government policy which would leave basic research unrestricted, but which would place some stringent controls on its application. No doubt the precise boundary between "basic" and "applied" science is sometimes hard to draw. But we have enough understanding of roughly where it comes to be able to use it as a basis for policy.

A. BASIC RESEARCH ON MAPPING AND SEQUENCING.

If we leave aside possible eugenic and other applications, the question about basic research is whether the knowledge it gives us will be harmful. It *is* likely to change our attitudes, in ways which have been indicated. Some of the possible changes are ones we may think undesirable: for instance, if a sharper awareness of inequalities between individuals or groups erodes equality of respect, or if the greater plausibility of causal determinism leads people in a confused way to a more fatalistic view of life. (It has been argued that clear thinking could avoid these particular changes, but perhaps we cannot rely on clear thinking being universal.) Other changes may be for the better: racist or sexist stereotypes might be undermined by the discovery that certain inequalities of genetic potential did not correlate with race or gender.

It is safe to predict that there will be changes of attitude, but the balance of benefit and harm is at this stage impossible to predict. This is partly because people may assess particular changes differently. But, more importantly, it is because we are talking about future discoveries. If we knew now what picture they will give, we would already have the knowledge and there would be no issue about whether to obtain it.

The unpredictability of specific changes of outlook resulting from this basic research might suggest that we have no rational basis for deciding whether or not it should be allowed to go ahead. Those of more cautious temperament might say that such unpredictability concerning matters so fundamental gives us a reason to stop the research.

The problem with this cautious view is that the argument supporting it applies to basic scientific research in general. All research carries with it the likelihood of unpredictable changes in our picture of the world, and so of unpredictable changes in our attitudes.

And there is a strong case against government intervention to protect our attitudes by stopping research which might undermine them. There are three main arguments:

1. To do so would be to opt for comforting ignorance or illusion rather than uncomfortable truth. The growth of science has rested on the preference for uncomfortable truth. Those of us who see science as one of mankind's finest creations will be dismayed by any wholesale repudiation of this preference.

2. It is unlikely that our attitudes can be protected by shutting down basic research. The knowledge that such protection was needed might in itself start to undermine them.

3. It may be that a government cannot stop basic research. It is not easy to monitor what goes on in laboratories. And what is stopped in one country may take place in another.

These considerations tell against trying to stop basic research on mapping and sequencing the human genome. They do not, of course, add up to a case for government funding. That case would need to be made in the usual way. The costs would need to be justified, in competition with other projects, either by the benefits of the eventual application of the knowledge, or by its pure intellectual interest.

B. APPLICATIONS: GENETIC ENGINEERING.

Because "gene therapy" carries with it a strong risk of unwanted side effects, it is too dangerous to use. Future developments may open up the possibility of safe use. If so, the government would presumably have a role in monitoring the safety of techniques, similar to its present role in monitoring drug safety.

What role should the government have if "positive" genetic engineering starts to seem a real possibility? The most defensible policy seems to be to start with a severely restrictive approach, possibly evolving into a second, more permissive phase.

1. THE FIRST PHASE: A MORATORIUM.

It is clear that the use of positive genetic engineering involves a real risk of major harm. There could be genetic disasters for individuals, perhaps ones passed on to future generations. There are also large unsolved problems about the limits of the genetic supermarket.

With these problems, there should be at least a legally backed moratorium, to give time for public discussion and for seeing whether any of the benefits are worth the possible risks.

2. A POSSIBLE SECOND PHASE: REGULATING THE SUPERMARKET.

Perhaps it would become clear in the public debate that no uses of positive genetic engineering would be acceptable. If so, there would be no second phase. But perhaps real benefits would be seen, together with solutions to the problems. If so, we might move cautiously towards a system where parents could make some genetic choices for their children. But, even then, it is hard to envisage that leaving the matter to unregulated market forces would be acceptable.

If the supermarket were to operate at all, we would need restrictions to protect people from harmful parental choices. We would also need restrictions based on the wider public interest. This would apply to such matters as the ratio between different genetic characteristics, as well as the need to preserve the variety of the gene pool. There might be restrictions to prevent economic inequalities turning into a gene-based caste system. In imposing these restrictions, the government would be acting as the custodian of the public interest. It would also be acting to protect future generations, who would also be affected by our blindly stumbling into unregulated positive genetic engineering.

BIBLIOGRAPHY.

PART ONE.

J. Passmore: *The Perfectibility of Man,* London, 1970.
P. Ramsey: Genetic Engineering, *Bulletin of the Atomic Scientists,* 1972.
A. Etzioni: *Genetic Fix,* New York, 1973.

A. Jones and W.F. Bodmer: *Our Future Inheritance, Choice or Chance,* Oxford, 1974.
J. Fletcher: *The Ethics of Genetic Control,* New York, 1974.
R. Nozick: *Anarchy, State and Utopia,* New York, 1974.
F. Hirsch: *Social Limits to Growth,* London, 1977.
President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research: *Splicing Life,* Washington, 1982.
J. Rifkin: *Algeny,* New York, 1983.
R.L. Sinsheimer: Genetic Engineering: Life as Plaything, in A.P. Iannone (ed.): *Contemporary Moral Controversies in Technology,* New York, 1987.

PART TWO.

G.E. Moore: *Ethics,* London, 1912.
H. Frankfurt: Freedom of the Will and the Concept of a Person, *Journal of Philosophy,* 1971.
G. Watson: Free Agency, *Journal of Philosophy,* 1977.
P.F. Strawson: Freedom and Resentment, in *Freedom and Resentment and other Essays,* London, 1980.

Mapping and Sequencing the Human Genome
Considerations from the History of Particle Accelerators

J.L. Heilbron (Berkeley) and D.J. Kevles (Pasadena)

1. Rationale and Organization

In December 1986, the Taskforce on Science Policy of the House Committee on Science and Technology issued an inventory of big-science instruments and facilities throughout the world. To qualify, an installation must have cost at least $25 million in 1984 dollars. Not one of the 275 items that appear is dedicated to biology.[1] A large federal project to map and sequence the human genome (MSHG) may well bring an entry from biology to the big-science list and a transformation of the fields concerned that some consider undesirable. Opponents worry about the centralized control and bureaucratization of big science and the possibility that a major investment in a large biology project would diminish the funds available for small-scale activities in other biological areas.

Professor Walter Gilbert of Harvard University, one of the most energetic proponents of MSHG, has called the human sequence "the grail of human genetics," which, like particle accelerators (the analogy may not hold here), would provide biologists with a magnificent tool for their most fundamental research.[2] Perhaps the strongest pressure for an immediate and substantial project in genome sequencing is fear of falling behind the Japanese, who are busily automating existing technology. Apart from the worry of scientists in the field that foreigners may preempt them, there is deep concern that whoever first perfects the technology of gene sequencing will achieve an advantage not only in information but also in manipulations essential to genetic engineering.[3] Commercial exploitation here intensifies international scientific rivalry.

Many commentators on MSHG besides Gilbert make reference to particle physics, alias high-energy physics, as the exemplar of the big science they deplore or promote. Both sides seem to have in mind big, contemporary, accelerator physics, as practiced in the United States primarily at the National Accelerator Laboratory (Fermilab) and the Stanford Linear Accelerator Laboratory (SLAC), both facilities of the Department of Energy (DOE), and abroad primarily at the European Laboratory for Particle Physics (CERN) and the Deutsches Elektronen-Synchrotron (DESY). The analogy so used, which assumes that MSHG will be conducted eventually at one or more federal centers, may assist considerations about organization, competition, cooperation, internationalization, and data access relevant to an eventual large, centralized effort. But we find that in its technology, institutional organization, and industrial relations, MSHG as presently practiced differs so much from contemporary high-energy physics that the usual analogy is not very illuminating. In many cases, much closer parallels, which bring out important considerations for policy toward MSHG, may be found not in the current but in the earliest regime in accelerator physics.[4]

High-energy physics began as little science in the 1930s. By then a gross physical map of the nucleus had been developed by chemists and physicists working alone or in small groups. They had discovered that the positive charge on the nucleus determines order in the periodic table of elements, that each element might have several isotopes different in atomic weight, and (after the discovery of the neutron in 1932) that the number of neutrons in the nucleus distinguished one isotope of an element from another. They had also provoked a few nuclear reactions using as projectiles particles from naturally occurring radioactive substances. Impatient with the skimpy streams of particles that nature provides, they built several different sorts of machines to acclerate rivers of particles to energies sufficient to penetrate nuclei. All the main accelerator designs of the 1930s are still in use for one purpose or another. But not equally in use. The machine of choice and prestige, which had come to dominate by 1940, was the cyclotron invented by Ernest Lawrence at the University of California and perfected at what became its Radiation Laboratory. The cyclotron brought large-scale, interdisciplinary science into academic environments. Its lineal descendants make the beams that serve the purposes of Fermilab and CERN. Machines like SLAC and DESY operate on principles introduced after the second world war.

In 1940, there were 23 cyclotrons in service or nearing completion in the United States, all financed primarily from private and State (not federal) sources. The largest, Berkeley's 60-inch, was being copied at the Carnegie Institution of Washington, and also in Tokyo and Birmingham, England. Berkeley forged ahead by starting to construct a cyclotron of 100 million electron volts. Less than eight years before, the achievement of one million volts had been regarded as a triumph. The hundred-fold increase in energy required parallel increases in staff and funding. One man, a graduate student in physics, who took the project for his Ph.D. thesis, built the million-volt cyclotron; a hundred physicists, engineers, and managers had their hands in the hundred-million volt machine. Money for the first came in small amounts and from traditional sources, around $1000 in all from the university's research funds and the National Academy of Sciences. Money for the second came in amounts almost without precedent and from an agency then only recently available to physicists, over a million dollars from the Rockefeller Foundation. By 1940, the Radiation Laboratory had ventured into, or invented, middling-sized science.

During World War II, the Laboratory grew wildly upon its basis of 1940. It entered big science, from which it did not withdraw after V-J day. It returned to the business of particle acceleration and nuclear science as a laboratory (now known as LBL) of the Atomic Energy Commission, which took over General Groves' nuclear empire on January 1, 1947. The AEC continued the Brookhaven National Laboratory (BNL), which Groves had founded in 1946 as an Eastern counterweight to the Berkeley Laboratory. In short order, accelerator physics became—and remains to this day—a ward of the federal government. The size and expense of its bigger machines put it beyond the means of individual universities and most sovereign nations. The Bevatron at LBL cost $10m in the early 1950s; SLAC, $100m in the early 1960s; Fermilab, $250m in the 1970s; and the Superconducting Super Collider, of uncertain future, perhaps $6000m in the 1990s. European nations shared the expense of big accelerator physics by founding CERN in the early 1950s.[5] Financial logic, if not perceived national and bureaucratic interest, recommends internationalization of the Collider and its successors in the 21st century. And there is another aspect of the wartime heritage that plays a prominent part in lobbying for, if not in creating, science policy. The physicists of Grove's Manhattan Engineer District learned the efficacy of crash prgrams. The tone of inappropriate urgency often sounded in their proposals may also be detected in pronouncements by some proponents of MSHG.

We do not suppose that key features of the history of particle accelerators will recur precisely in big-science biology. However, we believe that some analogies are close enough to shed light on how policymakers might wisely address several issues of importance to the genome project—its environment of technological development, scientific and practical returns, and accumulation and dissemination of map and sequence data. We have organized the balance of our analysis around these three topics. In each case we first state our presuppositions about MSHG and then give pertinent information about the history of accelerator physics. We end with an assessment suggested by consideration of this information.

2. Environment of Technological Development

The cost and time required to map and sequence the complete human genome will depend upon the quality and efficiency of the mapping and, especially, the sequencing technology. The federal government is providing support to encourage the development of this technology and improvements in it are now rapidly emerging from commercial firms and academic institutions. Advocates of MSHG take for granted that commercial interest secured through patents will figure prominently in the enterprise, but they also urge that the pace of progress be accelerated by the establishment of one or more federal centers devoted to MSHG.

Patents

The preponderant fact of economic life during the first phase of accelerator development, the 1930s, was the Great Depression, when support for basic scientific research plummeted. Under the circumstances, many academic scientists came to believe that professors and universities should seek to restore support by securing control through patents over inventions derived from their research. The University of California encouraged its faculty to work through the Research Corporation, a not-for-profit organization that aimed to secure patents for the inventions of academics and plow the proceeds back into university research. The Corporation tried to stimulate development of new industries based on patents it held while reinvesting its income in research deemed likely to further its industrial objectives.

Among the novel industries the Research Corporation had in mind in the 1930s was radiopharmacy. It supported Lawrence's research at Berkeley and he assigned it his rights to the cyclotron. The Corporation patented the machine, and tried unsuccessfully to patent useful radioelements made with it. Since, however, the cyclotron was by far the best machine for manufacturing radioelements, the Corporation believed that it could control the industry it hoped to create through its patent on the means of production. Although Lawrence at first doubted the propriety of protecting his invention, by the mid-1930s he spontaneously brought patentable inventions made in his laboratory and by physicists elsewhere to the Corporation.

However, neither he nor the Research Corporation tried to obtain royalties from the cyclotron. On the contrary, through the 1930s Lawrence opened his laboratory to would-be cyclotron designers from other institutions and he gave away radioelements for use in experiments throughout the world. (Had an industry developed around radioelements made at accelerators, as seemed likely in 1940, matters might have been different, but no such industry developed then because the invention of the radioactive pile during World War II created a far more powerful and economical producer of radioelements than the cyclotron.) The Research Corporation, which earned not a penny from its large investments in accelerators, accommodated gracefully. After the war, it wrote all cyclotron laboratories to grant royalty-free use of the machines. Many of the recipients of the letters had had no idea that a patent on the instrument existed. They had added their improvements to it—improvements in the source, vacuum system, insulation, oscillator, detector, etc.—without knowledge or worry that the basic design was protected.

This openness much accelerated the development of accelerators before the war. So did a similar policy pursued by the Atomic Energy Commission after it. Both law and policy have tended to vest in the AEC ownership of patentable inventions—for example, parts of accelerators—made in its laboratories or under its contracts and to make freely availabe the technologies of particle physics to scientists engaged in basic research. A similar attitude has governed particle detectors. The most recent achievement in that line, the Time Projection Chamber built at LBL and installed at

SLAC, has been adapted freely in accelerator laboratories in Europe and Japan.[6]

Institutions

During the 1930s, accelerators developed in a pluralistic, mainly academic institutional environment. Many important improvements in the cyclotron art were made outside Berkeley. Independence prompted doubts elsewhere in the country about the technical feasibility of the hundred-million volt accelerator. War interrupted its construction and also probably inspired, as it surely made possible, its completion according to a new design, which not only met, but far exceeded the energy goal of the old design. The episode shows two things: that Berkeley alone did not perfect accelerator technology and that a temporary redirection of effort may well inspire a novel technology overlooked or inaccessible in the initial rush.[7]

Institutional pluralism continued even after the federal takeover of particle physics. While the AEC invested in accelerators at BNL and LBL—and eventually at the Argonne National Laboratory, too—the agency also paid for similar if smaller machines at universities throughout the country. The system began to break down in the late 1960s, when the cost of new accelerators and the exhaustion of interest in physics at energies below the highest attainable forced decommissioning of old machines, and when attention shifted to fields with greater promise, smaller scale, and growing prestige, like solid-state or laser physics. Accelerators shut down everywhere: only clever maneuvering, refitting, and a switch to nuclear (from high-energy) physics and biology saved LBL's main accelerator, the Bevatron. In 1985, highest-energy machines again existed only at the national laboratories (BNL and Fermilab) and at SLAC, likewise a DOE facility. But contrary to the 1940s and 1950s, the middling-range "big-science" machines now concentrated in DOE laboratories too. In sum: as the machinery developed, research-level accelerators almost disappeared from the campuses while multiplying at the national laboratories.[8]

This concentration has not killed competition. LBL and BNL stimulated one another in the development of the first billion-volt accelerators, and BNL and FAL battled over the design of superconducting magnets for the current generation of highest-energy machines.[9] Still, the system is not set up to encourage innovation: it rewards engineering ingenuity, not radical new design. One of the arguments of the proponents of the superconducting super collider is that it will rely on proved technology. No doubt prudence has priority in a project estimated at several billion dollars. It was not the experts in superconducting magnets, the designers for FAL and BNL, who made the recent breakthrough in high-temperature superconductors.

3. Scientific and Practical Returns

All participants agree that MSHG requires the construction of a physical and genetic map of the genome. Much of the debate centers on the issue of pace and scope, particularly of the sequencing. On one side is a crash program that would emphasize the development as rapidly as possible of the requisite technology and then its use to sequence the genome starting, so to speak, at base pair number one and proceeding systematically to base pair number three billion. On the other side is a measured program that would encourage technological development for a period before attempting the whole genome and that would, along the way, use current technology to sequence those parts of it of clear scientific and medical interest. The argument for the former route is that the entire genome is worth knowing in sequential detail as rapidly as possible to support medical research into genetic diseases and biochemical study of protein manufacture, to forestall competitors, and to avoid the duplication unavoidable in a haphazard and decentralized mapping. The reasoning for the latter is that it is certainly worth knowing for certain chromosomal sites known to be interesting whatever mapping will reveal; but that the rest of it—estimated at 95% or more—holds too little of established genetic or biochemical interest to warrant setting all else aside to read it. [10]

Whatever the pace and scope of MSHG, it is widely assumed that the project will have a bounteous spinoff. Among them are prediction, control, and, ultimately, prevention of genetic diseases; an understanding of the basic mechanisms for the production of proteins and their actions; the stimulation of biotechnology; and the guidance of genetic engineering. Sequencing of the genomes of non-human species would make possible not only adventures in biological tinkering but also a profound account of the evolutionary interrelationships among animals. No doubt there will be much unanticipated cross-disciplinary fertilization as well.

Scientific Returns

During the 1930s, the Lawrence Berkeley Laboratory strove simultaneously to improve its machines and to advance nuclear science. The need to pay for the one and the wish to stay ahead in the other caused tensions never altogether resolved. To pay bills, the Laboratory explored the entire periodic table of the elements for substances that, when made radioactive, would find uses in medicine and in biological research. To conduct this exploration, to manufacture the desirable isotopes, and to improve the machines, Laboratory members organized into task forces that expedited the work but also restricted freedom of action.

The Laboratory used deuterons—the nuclei of heavy hydrogen—accelerated in its cyclotron as its main exploratory tool. (The deuterons activated targets directly or via neutrons they knocked from a piece of beryllium.) Much information accumulated about nuclear reactions. However, the quantitative results necessary to establish a reliable scale of nuclear masses or to investigate nuclear forces did not come from cyclotrons spinning deuterons but from other sorts of accelerators shooting protons. During the 1930s several of the most significant discoveries in nuclear physics—the neutron (made in England); the possibility of creating radioactive substances by bombarding ordinary materials with charged particles (France); and the peculiar effectiveness of the uncharged particle, the neutron, in provoking such transformations (Italy)—resulted from exploitation of naturally radioactive sources, not accelerators. Lawrence did not direct his cyclotron to splitting nuclei or to making radioelements until both had been accomplished elsewhere. The cyclotron put these discoveries within his reach instrumentally, since he soon repeated them with it, but outside his reach psychologically, since it had itself become the object of his researches.

During the 1930s, most of the time at accelerator laboratories did not go to research in physics or chemistry or biology, but to improving the machines and in using them in the routine preparation of radioisotopes for clinical or tracer work. Following up the discoveries abroad concerning artificial radioactivity, the Berkeley laboratory used the very strong beams procurable from the cyclotron to make radioelements, particularly isotopes of sodium and phosophorus and iodine, which appeared to have clinical applications. Searching the periodic table for new activities, the Berkeley Laboratory took on new regions of the table as soon as the cyclotron became powerful enough to excite nuclei there. The experimentalists chose (or were assigned) targets as they became available. Naturally much new was found, since no one had been there before. But the most spectacular discoveries came to people who concentrated their efforts on places they knew to be of particular interest. Three of these discoveries—technetium, K-electron capture, and the isobars of mass three—fell to men who followed their own paths through the Laboratory. The fourth discovery, carbon-14, was the result of a determined effort, pushed by Lawrence, to find a radioactive carbon isotope that could compete with the rare natural isotope, carbon-13, as a tracer in biological experiments. Carbon-14 made possible the untangling of the pathways in photosynthesis.

At Berkeley, the abler and more senior men resented the regimentation and policies that left less time for research than they wanted; and one of the objections raised against support for the hundred-million volt cyclotron in 1940 was that none of the grand experimental discoveries in nuclear physics in the 1930s had been made in an accelerator laboratory. After the war, the largest part of building and running accelerators became the province of engineers; physicists retired to perfecting and enlarging detectors, whence the engineers and computer scientists have followed them; with the consequence that all aspects of high-energy physics have fallen under the regime of large-team research.[11]

Practical Payoffs

Accelerator builders have long extrapolated to useful spin-off. It is notable that these claims and hints occurred more frequently in the 1930s, when performance had not constrained enthusiasm and investors had no independent reason to doubt the promises of cyclotroneers, than after the war, when the Office of Naval Research and the Atomic Energy Commission multiplied accelerator laboratories for their own purposes. More recently, particle physicists have thought it necessary again to argue their case with the help of anticipated spin-off, perhaps because their ever-increasing requirements no longer fit obviously among pressing national needs.

The spin-off indicated during the 1930s included power too cheap to meter and cures for cancer. A few hints from Lawrence that the era of practical atomic energy might be brought nigh by the cyclotron won him the enthusiastic support of the president of the University of California, who believed that his institution had an obligation to assist the State to solve its technical problems. Physicists divided over the practicality and imminence of useful atomic energy. The most sanguine were American. They could point to the very large production of neutrons by the cyclotron as an earnest of their projections: a neutron may enter a nucleus and transform it to another, with the release of energy; were it possible to release additional neutrons as well, a chain reaction might ensue.[12]

The discovery of a new sort of transformation—fission—that did release energy and neutrons came late in 1938. It did not take place in the United States or in an accelerator laboratory conducted by cooperative large-team work, or in an accelerator laboratory at all. It occurred in an institute in Berlin torn by Nazi politics; it was the work of two men, a chemist and a radiochemist, who had the advice by mail of the third member of their small team, a Jewish physicist in refuge in Sweden; and it required only primitive desk-top apparatus and radioisotopes made by perfectly standard methods.[13] Accelerator laboratories confirmed and extended the discovery of fission and, eventually, did the preliminary work toward the manufacture of a uranium bomb (by perfecting the electromagnetic method of isotope separation at the magnet for the hundred-million volt cyclotron) and a plutonium bomb (by preparing weighable samples of plutonium in the "medical cyclotron" completed there in 1939).

The medical cyclotron was built partly to furnish radioisotopes for experimental and clinical use and partly to deliver a beam of neutrons to burn out cancers. It appeared for a time that neutrons did more damage to tumors and less to healthy tissues than the "equivalent" exposure to high-voltage x rays. Treatments at first gave encouragement, but proved disappointing in the long run. Here again, however, the application, though premature, presaged something much more effective; and now particle-beam therapy has a place in the war against cancer.

In sum: particle accelerators, conceived originally for research in nuclear physics, and recommended, by some, particularly in the United States, as keys to unlock "the cosmic cupboard of energy," came to have spin-off of enormous importance owing to discoveries made with much simpler apparatus by physicists and chemists in several European countries. Although they have not delivered electricity too cheap to meter or a cure for cancer, cyclotroneers of the 1930s did help to bring in nuclear explosives, nuclear power, a radiopharmacopeia as well as diagnostic aids, and radioactive tracers that have made possible elucidation of complicated biochemical processes.

Spin-off from accelerator laboratories in the postwar has been of a different character and importance from that of the earlier period. The decision to proceed with a redundant investment in accelerator laboratories at both Berkeley and Brookhaven related to worry about manpower. Building competing machines at LBL and BNL would keep Berkeley's group of practiced engineers and physicists intact and create a parallel one in the East for service in a new Manhattan District, should the Soviets make one necessary. It was also recognized that, together with LBL and BNL, the academic accelerator laboratories would train new physicists and engineers who might—indeed did, in large numbers—staff the national laboratories and the facilities of the AEC's contractors. Then, too, the Cold War was taken to dictate that the United States be preeminent in particle physics—a claim advanced nowadays by advocates of the Supercollider. In all, competition internationally and competition within the system as well as concern to create skilled cadres brought Nobel prizes, national prestige, substantive knowledge, and a large number of Ph.D.s in science and engineering.

Another sort of spin-off has been encouragement of satellite industries. The pattern adumbrated in the 1930s, when what became Litton Industries served the Radiation Laboratory at Berkeley and Varian Associates got its start around an electron accelerator at Stanford, firmed and multiplied after the war. Data processing requirements have stimulated the development of specialized computer software, much of which, however, seems to be done in-house. Perhaps the most useful direct technological spin-off from accelerator building after the war is the synchrotron source, a plentiful supply of high-energy x rays employed in, among other things, industrial radiography. Berkeley's old medical cyclotron continuesd its humanitarian service. It has been adapted to the analysis of the chemical composition of rare antiques, the constitution of the ink in Gutenberg's bible, and frauds and counterfeits among genuine objets d'art.

4. Databases

The output of the sequencers will be stored in an electronic database. Following the precedent established by the AEC, the DOE has manifested an interest in the health effects of nuclear radiation, particularly the impact of radiation on plant and animal genomes. It has joined with the National Institutes of Health (NIH) to create GenBank, a computerized inventory of base sequences (currently numbering perhaps 10 million). DOE maintains GenBank at its laboratories in Los Alamos LASL and Livermore LLNL. DOE has also established a National Human Gene Library consisting of bits of chromosomes, copies of which have been distributed around the world since 1985.[14]

Neither the formation nor the operation of the MSHG database is likely to be simple or straightforward. In June 1986, Genbank was about two years behind in entering sequenced data and had on deposit only a fifth of the sequences published in 1985. NIH has fostered the development of software for accessing sequence data but the existing network—GENET—was so overwhelmed in its first six months that it was compelled to deny commercial users further access. It appears that GenBank may not have recieved adequate cooperation from depositors and users. GenBank has had to do most of its own evaluation since only 30% of sequencers requested to annotate their findings did so. This annotating cost 30 cents/base for the first six million bases—a cost that may not be so amenable to reduction through improvements in automation as the sequencing itself.[15] GenBank has had to omit signalling gene boundaries, borders between introns and exons, etc., in order to cope with its backlog. Then, too, potential contributors to GenBank and GENET worried that their right (as they saw it) to first analysis of their data might be compromised if they allowed its incorporation before they had finished publishing or exploiting it.[16] The MSHG database poses problems of construction, access, and control.

Encouragement from Physics

Database practices in high-energy and nuclear physics suggest that some of these problems may be tractable. The neutron data bank maintained by the Nuclear Energy Agency (NEA), a service of OECD, and the Evaluated Nuclear Structure Data File (ENSDF) at Oak Ridge National Laboratory are relevant and representative. The NEA's bank has four copies, one at Brookhaven, which has responsibility for the main file of evaluated neutron data. ENSDF, which contains data about isotopes, also belongs to an international network coordinated by the International Atomic Energy Authority. A world-wide wish list called WRENDA, constructed from users' requests for unavailable or non-existent data, helps to guide further measurements to places where it is wanted.[17] The wishers have to wait. BNL aims at a cycle of six years from submission of new data to its evaluation and incorporation in the neutron bank. ONRL strives for four years to process new information about isotopes.

Certain physics databases conducted under rules established by the International Atomic Energy Authority have an unevaluated and an evaluated part, both available; the objective is to make the data flow from the one part to the other through committees of experts. Other physics databases are never evaluated. The high- energy physicists can do without expert assaying, perhaps because they are all experts and their data, which has little application, are of interest only to themselves. They hurry to obtain new data in higher energy regions, not to improve old data in lower ones. Their wish list consists of bigger accelerators.[18]

The apparently smooth international cooperation in the maintenance and dissemination of at least some large databases in physics may presage the possibility of similar success with sequence data. The cost and turn-around time for evaluated information are also useful if not happy precedents. Another feature deserves mention: much of the consultation of the physics databases has occurred via experts, usually members of the staff at BNL, ORNL, OECD. It is not a good assumption that the consumer of MSHG data will be able to obtain it directly from an on-line computer hookup, especially if he or she must search several databases constructed on different principles.[19]

5. Assessment

Although a full-scale project in MSHG would have something in common with a modern, national, accelerator laboratory in cost and perhaps also in organization, the current state of MSHG suggests more compelling analogies to the accelerator practice and nuclear science of the 1930s. The state of the mapping of the genome compares with understanding of the gross features of the nucleus secured between 1910 and 1930; the state of sequencing technology compares with the adventuresome designing of accelerators in the 1930s. There are also strong negative analogies. MSHG will demand a degree of coordination in the construction and management of its database that nuclear and particle physics have never faced.

Technology

The similarity to knowledge about the nucleus before 1930 lies in the fact that a rough physical and genetic map of the human genome remains to be constructed. The similarity to accelerator technology in the 1930s lies in the fact that all phases of the technologies of MSHG are rapidly developing, often in competing forms, at many public and private institutions. New sequencing machinery is highly likely to reduce the cost and the rate of sequencing by one or more orders of magnitude.[20] No single technology has emerged so dominant over all others and so complicated and expensive as to compel its development in one or several major federal laboratories. A key dissimilarity to accelerator physics up to the present day is the degree to which technological innovation in MSHG results from commercial activity. At a recent conference at Cold Spring Harbor on human molecular biology, some 25% of the papers came from the corporate sector. Commercial restrictions have been interfering significantly with the free exchange of results and ideas. Similar problems could well invade academic efforts in MSHG, since federal policy now encourages investigators in non-profit institutions to collaborate with commercial firms and permits non-profit recipients of federal research grants to patent inventions made in their laboratories. Two reviewers of progress and prospects in DNA sequencing have noted that "a disproportionate amount of the available literature is in the form of patent documentation."[21]

The technological similarity to accelerators in the 1930s would seem to work against the establishment in the immediate future of major federal MSHG centers for technological development. To be sure, the Japanese propose a single national center—to be efficient, it would need to be continuously fed with genetic samples from around the globe—for the near future using the technology they are developing with Seiko and Fuji Photo.[22] A similar American commitment now would be likely to freeze the relevant technology prematurely. Further, a facility devoted to technological development in conjunction with sequencing runs the risk of falling into the same trap that ensnared Ernest Lawrence in the 1930s—a focus on machinery that could prove costly in the coin of missed scientific opportunities. The reasons—or rather the rhetoric—for haste do not seem compelling. Proponents of a crash program have not shown that the genome gap of 1987 is any wider than the misile gap of 1960.

Since the state of MSHG technology so closely resembles that of particle acclerators in the 1930s, federal policy would do well to foster what worked so well in the physics case—pluralism, decentralization, and local initiative. The principal mechanism is obvious: a program of grants under peer review to individual or small groups of investigators to develop the physical and genetic map of the genome and to forge improved and lower-cost technologies of mapping and sequencing. But special consideration must be given to the dissimilarity to accelerator projects in the 1930s—the chilling effect that patentability seems to be having on the circulation of MSHG ideas and information. The statutes governing property rights in processes and inventions produced by work under federal research grants furnish the granting agencies with mechanisms to ensure the dissemination and circulation of results. These mechanisms should be used to achieve open and efficient technological development in MSHG.[23]

If the equipment for sequencing proves to be small, inexpensive, and reliable, then MSHG could continue to operate primarily through small-scale installatons. The cyclotron was first recommended for just such virtues, in comparison with high-tension accelerators; but as we know it soon outgrew its competitors. (Scale-up is not mandated by nature. The Tiselius electrophoresis machine of the 1930s was as large and expensive as a mid-size cyclotron; it developed in the opposite direction, and now electrophoresis, which is an essential step in all the sequencing schemes under discussion, can be done in inexpensive table-top instruments.) There are dangers in this decentralization. The presence of a local capacity for sequencing could bring into the research laboratory technical, unimaginative, repetitive operations that have no place there. The use of university cyclotrons to manufacture radioisotopes for medical use and of university engineering laboratories for routine testing for industrial laboratories shows the forces at work. On the brighter side, a multitude of cheap sequencers would decentralize control, provide room for independent projects and approaches, allow easy corroboration of results, and promote still further technological innovation.

Should the equipment for automating the sequencing and interpreting the results be bulky, costly, and temperamental, a situation similar to that in particle physics may result: the number of productive centers will diminish and innovation will become still more costly until, perhaps, only a single federal sequencing installation remains. Much may be said for a centralized, routinized, laboratory—for example, economy of scale, consistency, and quality control. And, no doubt, work on a super sequencer would go on in the neighborhood, and on the advice, of programs in genetics and biochemistry. However, large facilities devoted to gene sequencing would no doubt suffer the same drawbacks as those involved with big physics, but without the same sort of payoff. Cartography does not win Nobel prizes. As the leader of the Genetic Group at Los Alamos observed, those who have worked on GenBank have not published the half-dozen papers a year customary for energetic researchers in biology. The hole in their bibliographies would not assist advancement in an academic environment.[24] The social consequences of big science are known: cooperation, interdependence, homogeneity among the lower echelons; entrepreneurism and dirigisme among the principal investigators. If genome sequencing occurs in such a setting, we might expect teams of technicians, graduate students, and post-docs overseen by senior investigators who plan the work and utilize its product. Not all creative physicists have had the temperament for team research. Some very distnguished ones have left the field out of distaste for it. There is no reason to expect that big biology would fare differently.

The question of the disposition of any such major federal laboratories after the sequencing of the human genome is completed should be faced at the outset. Precedents from the record of the AEC laboratories suggest the need for a clear eye. Initially LBL and BNL were the only "multi-purpose" members of the system; they vied where they duplicated one another's efforts, in high-energy physics. The AEC operated other laboratories too: Argonne, Los Alamos (LASL), Oak Ridge, and Pacific Northwest, which trace their origins to the war; Sandia, which developed from the engineering and assembly division of LASL; and Livermore, which developed from a unit of LBL set up to compete with LASL. The missions of these laboratories were initially limited to specific purposes: reactors, isotope manufacture, weaponry, etc. The commissioners of the AEC soon learned, however, that they could not maintain morale or retain technical personnel unless they offered inducements in the way of research. Thus the missions of the laboratories broadened to include fields remote from immediate application, not least because of their interest in sustaining themselves. The record shows that it is possible to kill a program at a national laboratory, but not the laboratory itself.[25]

Finally, policy makers must consider the role of the national laboratories in promoting what is likely to be a desirable feature of MSHG—international cooperation. High-energy and nuclear physics have long been cosmopolitan: its practitioners travel from one installation to another, experiment here and there, meet at frequent international conferences, etc. CERN outside Geneva, Switzerland, is the exemplar of international cooperation in big science: it belongs to the several European governments who pay its bills, it has an international board and staff, it admits scientists from all over the world. There are also important bilateral laboratories, for example, the Laue-Langevin, German- French, high-flux reactor in Grenoble. DOE laboratories are not international in this sense. They do indeed admit foreign scientists to do experiments. But all U.S. facilities in high-energy and nuclear physics are owned, operationally funded, and managed by American organizations and personnel, and Americans are by far the most frequent users of the machines.[26] Although DOE has a great man cooperative programs in place with many nations—some 97 agreements with 25 countries at the end of 1984—these arrangements do not involve the laboratories very deeply. Most of DOE's "international" activity is reconnaisance—exchange of information, samples, instruments, specialists, visits—and aims at gaining access to foreign technology, data, information, and markets.[27]

Until quite recently, American high-energy and nuclear physicists possessed the most powerful machines for their purposes in the world. They cooperated from positions of unchallenged strength. The ancien régime might be illustrated by LBL's and BNL's help to CERN during the construction of its first proton synchrotron, by American supply of radioisotopes throughout the world during the 1950s, and by the dominance by Americans of international meetings on accelerator technology through the 1970s. Now CERN threatens to jump ahead. It has near completion an accelerator that will reach higher energies than any in existence or under construction in the United States. The pressures for an American machine—the Supercollider—to outdo the Europeans come largely from the national laboratories.[28] They do not favor a truly international facility. Insofar as the past policies of the national laboratories anticipate the future, their internal competitions, efforts to diversify, and experience of world leadership do not suggest that they would push vigorously to internationalize the sequencing of the genome.

Whether organized on a large or small scale, MSHG seems unlikely to achieve significant international cooperation in the development of sequencing technology. The perceived stakes in commercial and international rivalry are too high.

Database

The best parallel between the likely future of MSHG and the current state of big physics concerns not sequencing technology but the data base—the physical and genetic map of the human genome and the sequence of base pairs. Here is perhaps the mos fundamental tool for future researchers in medicine and biology.[29] It compares in importance and centrality not with the reference data of nuclear and particle physics but with their great instruments of research. The MSHG data base must therefore be accessible, reliable, and usable. The construction and management of such an extensive, high-quality database will be a major enterprise.

As we have noted, the cost of entering and annotating the data is high and not all creators of sequence information like to cooperate in the construction of the analyzed database. Commercial institutions may wish to withhold sequence data from the database entirely for proprietary reasons: serious attention is now being given to the possibility of protecting sequence data under copyright.[30] Researchers in non-profit institutions and universities may wish to keep the data back for their own use. To be sure, the right of prior use has been limited by court decisions that expose raw data collected under federal grants to calls under the Freedom of Information Act insofar as the data may be considered agency records. Still, there seems to be no firm policy or principles governing confidentiality or copyrightability of the results of federally-funded research, including genetic maps, sequences, and databases.[31]

The federal government has the resources to pay for data. (Its willingness to do so might be used as an incentive for commercial firms to perfect mapping and sequencing technology and also, in view of the great uncertainty about the eligibility of sequence data for copyright, to provide such data in the extensive areas of the genome unlikely to attract medical or scientific interest.[32]) It may also have the power to compel its acquisition from those whose research it supports. Because of the immediate relevance of sequencing to genetic engineering, and because of the potential national and international importance of its products, the government might impose restrictions on the ownership of results of sequencing done under federal grants.[33] As the National Science Foundation declared in 1982, it is imperative that policy regarding the production, dissemination, and use of scientific and technical information be an integral part of R&D policy.[34] In the case of MSHG, federal granting agencies might compel grant and contract recipients to submit their map and sequence data to the database on a standardized, annotated form. Funds might also be made available to purchase map and sequence data from profit-making biotechnological firms.

To construct and manage the database, some federal entity might be required at an early date.[35] Mapping needs to proceed apace and the physical map will yield on the order of 100,000 cosmids—a first-order challenge to database construction and management. The entity would coordinate mapping and sequencing programs, develop standards for contributions to it, see to their reliability, and arrange for the sequencing—possibly by contract to outside institutions—of the seemingly less interesting parts of the genome. These functions would demand an R&D capability. It would be in order, therefore, to establish in conjunction with this entity a laboratory—a National Bureau of Biotechnical Standards,say—the principal duties of which would be to provide technical support for the development and management of the database, including checking its reliability and operation and seeking ways to improve its use and accessibility. Should sequencing technology develop in a way that requires its installation in a federal center, it could be installed and operated in the NBBS. This would be a secondary activity for the Bureau and it could return entirely to its primary duties after the sequencing of the human genome.

The proposed entity would also be responsible for international cooperation in MSHG. Cooperation may not be significant in the technological area, but may well be so in the informational one. With regard to basic scientific information, which any country can produce ultimately, the precedents of particle physics suggest that foreign countries, even those that would compete with us, would be inclined to share and collaborate. Of all subjects susceptible to international cooperation, sequencing the genome of the human species, which is not peculiar to any race or region, is perhaps the most attractive. It might lend itself better to engendering international good will and intercultural exchange than experiments at big accelerators.

6. Summary

MSHG has multiple goals: to foster the physical and genetic mapping of the human genome; to stimulate the ongoing developmnt of sequencing technology; to generate, eventually, a sequence of the complete human genome. Our reading of the relevant history of particle accelerators suggests that for the next five to ten years it would be wise to pursue these several goals by mixed and multiple means designed to take advantage of the pluralist forces of scientific and technological ambition already manifest in the field. To wit:

1) a loosely coordinated program of federal grants, awarded under peer review, to investigators at non-profit institutions for the purpose of encouraging work on mapping and the development of sequencing technology, subject to federal regulations applicable to whatever patentable technologies may be developed under such grants;

2) the establishment of a federally owned and managed map and sequence data base, to which federal grantees would be required to submit map and sequence data in standardized, annotated form;

3) the creation of a suitable entity to oversee the data base, with appropriate technical support from, perhaps, a National Bureau of Biotechnical Standards; this NBBS to have the authority to buy sequence data from research institutions without federal grants and to contract for the production of such data; and to undertake sequencing on its own as a secondary activity if and when the trend of the technology seems to warrant it.

Notes

1. U.S. Congress, 99th, 2nd session, House, Committee on Science and Technology, Task Force on Science Policy, *World inventory of "big science" instruments and facilities* (Washington: GPO, 1986).

2. *Science*, 27 June 1986, pp. 1598, 1600; *Harvard Gazette*, 20 Feb 1987, p.5; Liebe F. Cavalieri, "Gene Sequencing: No Easy Answers," *The scientist*, 26 Jan 1987, p. 13; David Baltimore, "Genome-Sequencing: A Small-Science Approach," *Issues in science and technology*, spring 1987, pp. 48-50.

3. Lee Hood, *Insight*, 11 May 1987; Health and Environmental Research Advisory Committee (HERAC), Subcommittee on Human Genome, "Report on the Human Genome Initiative," draft, March 1987. Two Japanese investigators have predicted that the automation of complex laboratory procedures "could well turn out to be the equivalent of the Industrial Revolution in biological and biochemical laboratories." Akiyoshi Wada and Eiichi Soeda, "Strategy for building an automatic and high speed DNA-sequencing system," manuscript, 1986, p. 10.

4. All the information about the early history of particle physics in this report is documented in J.L. Heilbron and R.W. Seidel, *A history of the Lawrence Berkeley Laboratory*, vol. 1: *From the beginnings to Pearl Harbor* (Berkeley: University of California Press, in press).

5. A. Hermann, J. Krige, U. Mersits, and D. Pestre, *History of CERN*, vol. 1: *Launching the European Organization for Nuclear Research* (Amsterdam: North Holland, 1987).

6. DOE, *Annual report*, 1983, p. 61.

7. The design of the Bevatron profited from the same sort of delay. J.L. Heilbron, R.W. Seidel, and B.R. Wheaton, *Lawrence and his laboratory: Nuclear science at Berkeley, 1931–1961* (Berkeley: Office for History of Science and Technology, 1981), pp. 62–79.

8. Ref.1, pp. 76–88.

9. Lillian Hoddeson, *Historical studies in the physical sciences*, in press.

10. Francisco J. Ayala, "Two frontiers of human biology: What the sequence won't tell us," *Issues in science and technology*, spring 1987, p. 56; *Science*, 27 June 1986, p. 1600.

11. Peter Galison, *How experiments end*, in press.

12. Lawrence Badash, Elizabeth Hodes, and Adolph Tiddens, "Nuclear fission: Reaction to the discovery in 1939," American Philosophical Society, *Proceedings*, *130* (1986), 196–231.

13. Fritz Krafft, "Internal and external conditions for the discovery of nuclear fission by the Berlin team," in W.R. Shea, ed., *Otto Hahn and the rise of nuclear physics* (Dordrecht: Reidel, 1983), pp. 135–65.

14. Roger Lewin, "Proposal to sequence the human genome stirs debate," *Science*, 27 June 1986, pp. 1598–1600.

15. Ibid., p. 1599.

16. Roger Lewin, "National networks for molecular biologists," *Science*, 30 Mar 1984, pp. 1379–80; cf. Lewin, "Long-awaited decision on DNA database," *Science*, 27 Aug 1982, pp. 817–18.

17. F.D. Gault, "Physics databases and their use," *Computer physics communications*, *22* (1981), pp. 125–32.

18. Ibid., pp. 126, 130–1.

19. Ibid., p. 131.

20. William J. Martin and R. Wayne Davies, "Automated DNA sequencing: Progress and prospects," *Biotechnology*, 4 Oct 1986), pp. 890–4; *Science*, 4 July 1986, p. 24.

21. Statement of Susan C. Rosenfeld, Science and Law Committee, Association of the Bar of the City of New York, at the Office of Technology Assessment

Workship on Issues of Collaboration for Human Genome Projects, 26 June 1987, pp. 6-7; Martin and Davies (ref. 20), pp. 894–5; *Science*, 11 July 1986, p. 157.

22. Wada and Soeda (ref. 3), pp. 10, 12. The Subcommittee on the Human Genome of DOE's Health and Environmental Research Advisory Committee recently observed, "We think it will be more effective to develop new and original procedures....Flexibility and innovation should be fostered. It is particularly important in this rapidly developing field not to start any large inflexible organizations whose direction would be hard to change."

23. Rosenfeld (ref. 21), pp. 8-10.

24. Robert K. Mayzis, interview in *Los Alamos science*, spring/summer, 1985, pp. 90, 91.

25. R.W. Seidel, "A home for big science: The Atomic Energy Commission's laboratory system," *Historical studies in the physical sciences, 16* (1986), 135–75; Albert H. Teich and Henry Lambright, "The redirection of a large national laboratory," *Minerva, 14* (1976/7), 447–75.

26. Ref. 1, pp. 6, 11, 45.

27. DOE, *Annual report*, 1985, pp. 151, 153.

28. E.g., Sheldon L. Glashow and Leon M. Lederman, "The SSC: A machine for the nineties," *Physics today, 38* (Mar 1985), 29–37.

29. Leroy Hood and Lloyd Smith, "Genome sequencing: How to proceed," *Issues in science and technology*, spring 1987, p. 38; Walter Gilbert, "Genome sequencing...," ibid., p.34.

30. Rosenfeld (ref. 21), pp. 12-13.

31. Dorothy Nelkin, *Science*, 14 May 1982, pp. 704–6; National Science Foundation, *Emerging issues in science and technology* (Washington: NSF, 1982), p. 83; cf. U.S. Congress, OTA, *Commercial biotechnology: An international analysis (Washington: GPO, 1984), OTA-BA-218.*

32. This would be an alternative to establishing a few large centers to sequence the 95% or so of the genome that is assumed to be of less than compelling interest; Rosenfeld (ref. 21), p. 12. For the suggestion of centers, see Hood and Smith (ref. 29), p. 43.

33. HERAC's Subcommittee on the Genome has insisted that "all new map and sequence information must be placed promptly in a designated data base." Ref. 2, pp. 2,3.

34. National Science Foundation (ref. 31), pp. 73, 79, 81–2.

35. Many biologists hold that coordination of emerging DNA databases should be a top priority. *Nature*, 23 Oct 1986, p. 660.

Horace Freeland Judson
Wissenschaftskolleg zu Berlin
Wallotstrasse 19
D-1000 Berlin 33
Federal Republic of Germany

phone: 011-49 30 89-001229

For The Office of Technology Assessment REVISED VERSION 28 September 1987

Mapping the Human Genome: Historical Background

by Horace Freeland Judson

Genetics--the explanation of how organisms pass on traits of anatomy, physiology, and behavior to their descendents and of how each individual expresses those traits in its formation and throughout its life--is the central problem of biology. Which is to say, as we gain fuller understanding of the transmission and expression of hereditary characters we progressively unlock understanding of whole classes of other problems in biology, as well. These include problems both basic and applied (if, indeed, the distinction retains any utility). Thus, genetics underlies all of the biology of cells, including developmental biology, or embryology, but also enzymology and the study of cancers and many other diseases; all of immunology and of endocrinology; neurobiology generally as well as many disorders of the nervous system and of the mind; and, ultimately, the processes of evolution. Though no one supposes that genetic analysis will be sufficient to explain everything we want to know in these vast scientific domains, every biologist recognizes that genetics is necessary to all such explanation. The necessity is not only logical but methodological--for genetics is the crucial mode of access to the rest.

Central to genetics, of course, is the concept of the gene: but genes act only in concert, which is to say that the concept of the gene includes the map of genes and their sequences. Modern genetics began in 1900, with the rediscovery of Mendel's rules and Mendel's paper of thirty-five years earlier. The word gene first appeared in 1909. The first demonstration that a particular gene had a locus--that it could be assigned to a particular chromosome--was published in 1910; the first genetic map, which showed the relative locations of six genes on one chromosome, appeared in 1913. As the experimental systems of organisms and techniques that geneticists employ have grown more sophisticated, the concept of the gene has changed and deepened--and with it, integrally, the idea of gene maps and sequences. In 1976 and 1977, three-quarters of a century after the rediscovery of Mendel, the first total genetic sequences (what that entails we develop below) were determined, in each instance for a microorganism with fewer than ten genes all told. Now, a decade later still, some biologists propose to sequence the entire gene set of a creature as complex as any--man--whose genes number at least fifty thousand and perhaps several times that. What this proposal signifies, how it can be carried out, what gains it could bring--all these emerge from the historical development of genetics from 1900 to the present day. And although genetics today displays a rococo intricacy of detail, the main line of its history is straightforward and casts a brilliant light upon the latest methods and discoveries.

Two preliminary observations will clarify that main line. First, the fundamental discoveries of genetics emerged as a series of approximations. One can think of these as a succession of models of the processes of heredity; they were also successive redefinitions of the gene and of the nature of gene maps and gene sequences. Each model, each redefinition, in order to incorporate more of the evident diversity of inheritance, refined and made more complex the version preceding it. Technically, too, genetic analysis has proved surprisingly conservative: even today we see the earliest methods of mapping and sequencing not superseded but adapted and readapted.

Second, heredity necessarily has that dual aspect, as the transmission of characters from one generation on to the following and as the expression of characters in the developmental process by which an individual organism builds itself. From the start, new discoveries typically related more to one side or the other of that duality. The transmission of characters and their expression were triumphantly united, however, in the discovery, in 1953, of the molecular structure of the genetic substance--deoxyribonucleic acid, DNA, the celebrated double helix. Today, the duality takes on ethical as well as technical importance: we await the imminent first trials of gene therapy to correct the expression in individual patients of certain inherited human disorders, even while we contrast such prospective cures with the more remote but also more troubling idea of genetic alterations that would be inherited by an individual's offspring.[1]

From Mendel to the first genetic map

Charles Darwin desperately needed an adequate explanation of heredity--of the appearance of heritable variations in populations--to complete the mechanism of his theory of evolution. He spent years elaborating such an explanation, and published it in 1868 (nine years after the first edition of The Origin of Species) in two big volumes. But his hypothesis failed, in large part because he tried to encompass all the diverse phenomena of heredity that his extensive observation and even more extensive correspondence and reading had made him aware of.[2] Gregor Mendel published his experiments on plant hybridization in 1865, in a paper of fifty-five pages. He wrote by way of introduction that a full explanation of heredity would shed light on evolution. He started by offering the most elementary model.

Mendel, to be sure, did not think in terms of maps or sequences somewhere within the cell, but rather in terms of the statistics of inheritance of visible characters. His genius was to find in his pea plants certain discrete pairs of alternative characters--round seeds versus wrinkled ones, tall plants versus short, and so on, seven pairs in all--that were inherited in a remarkably simple way. He traced these pairs through successive generations of hybrids. In the first place, of any pair each plant displayed either one trait or the other: the alternatives were non-blending. (For example, his pea plants were either tall or short, never of medium height.) Second, every pair was inherited independently of the other six pairs: the characters were free-assorting. (For example, a tall plant could have round seeds or wrinkled seeds, as could a short plant, and so on with the seven character-pairs in any combination.) Third, a trait, one of a pair, could persist without being exhibited through one or many generations and eventually reappear: the latent trait, Mendel said, was recessive and the visible character dominating--terms that, of course, we still use. (For example, when he crossed a tall plant from a pure-bred line with a short plant from a pure-bred line, he got only tall plants in the first generation. Tall was dominant. But when he crossed those hybrids, about a quarter of the next generation were short.) All this led him to see that each parent plant contributed one of a pair of elements, and no more than one, to determining each character in the makeup of each offspring plant. If an offspring was hybrid for a given character, one of those elements but not both being recessive, then in that character the plant would look like the dominating form but--an essential point--would pass the elements to its own offspring at random, about half dominant and half recessive.

These observations now seem simple and familiar. They were revolutionary. With them, Mendel established many of the fundamentals of genetics. We must analyze the heredity of an organism, at least in the first place, by its various unit traits. These are determined by elements that typically occur in pairs, with one of each pair received from each parent. The elements of a pair don't blend, and, further, one of a pair may be masked by the other without being lost. In determining the next generation, each pair is split up and distributed at random. Therefore, one must think about heredity statistically, looking at relatively large numbers of offspring in successive generations, to discern the patterns.[3]

The reasons Mendel's revolution hung fire for three decades are many and not obvious. Mendel was not the obscure monk of the vulgar myth but a scientist trained at the University of Vienna and working at a monastery of a teaching order in an important provincial city of the Austrian empire, where he later was elected abbot. After he published his paper, a number of copies circulated in Europe, even in England. (One may have been available to Darwin, but we can be confident that he would not have understood it--although his cousin Francis Galton, who thought about heredity deeply and statistically, would have seen the point at once.[4]) But Mendel's style of analysis was altogether unfamiliar to biologists--those unit characters, non-blending, randomly assorting, to be analyzed statistically. Statistics was then in its infancy. More important (and more impressive in relation to what he discovered), Mendel knew of no physiological machinery, no anatomy or processes within the cell, by which to explain the abstract algebra of his hereditary elements.

Between 1865 and 1900, statistics developed enormously. So did cell biology, or cytology, a sister science that provides genetics its anatomical basis. Biologists made the first close observations, under the microscope, of the sperm cell entering and fusing with the ovum, and of what happens when cells reproduce. They noted that cells have nuclei. In 1877, they saw that thread-like structures they termed chromosomes appear in the nucleus as the cell is about to divide. Microscopists teased apart an elaborate series of steps by which, in cell division, the chromosomes align themselves in matched pairs, then double and split up, complete and apparently identical sets going to each daughter cell. They distinguished a second kind of cell division--a more elaborate dance of the chromosomes still--whereby the germ cells form themselves, each sperm or ovum containing a half set of chromosomes, exactly one of each pair. Once that halving process was recognized, its point was obvious. At the moment of fertilization the new individual reconstitutes a full complement of chromosomes with one of each pair from each parent.

In 1892, the noted German physiologist August Weismann put all this down in a book, Das Keimplasma (The Germ-Plasm), which included the drawings of the stages of cell and chromosome reproduction that have plagued every high-school biology student since. High-school stuff, perhaps--but Weismann was able to say that at last it was settled that the male and the female parent contribute equally to the heredity of the offspring; that sexual reproduction thus generates new combinations of hereditary elements; and that the chromosomes must be the bearers of heredity. What the chromosomes bore were ultimate particles, determinants for each separate character of the organism.

Acknowledging that the exercise was sheerest speculation, Weismann estimated the number of these determinants in terms of the number of characters to be accounted for and of the sizes of molecules as chemists knew them. He needed perhaps millions of determinants and worried that they wouldn't fit in the chromosomes. Speculation, yes--but Weismann was first to think seriously about the hereditary material as a physical substance, and its necessary characteristics. Weismann was read by biologists everywhere; his books were translated promptly into French and English, The Germ-Plasm appearing in 1893, dedicated to the memory of Charles Darwin.[5]

Thus in 1900 when Mendel's rules and his paper were rediscovered, independently by several botanists, the community of biologists was prepared to consider heredity cytologically and statistically. Developments came fast. Mendel was translated into English at once, by William Bateson, an English botanist. In 1902, an English physician, Archibald Garrod, reported in The Lancet his observations of a disorder, alkaptonuria--in which the patient's urine on exposure to air turns terrifyingly but harmlessly black--which he had traced through several generations in several families. It was inherited, he said, in a Mendelian fashion. It appeared to be a blockage of one step in a metabolic pathway, that is, a specific biochemical reaction. He called alkaptonuria an inborn error of metabolism. Garrod was first to demonstrate Mendelian inheritance of a character in man. He was first, also, to connect Mendelian heredity with the biochemical pathways of the individual organism. He identified several other inborn errors of metabolism, including cystinuria, which shows itself as a tendency to form bladder stones, and albinism.[6] The latest edition of Mendelian Inheritance in Man, the seventh, edited by Victor McKusick, catalogues nearly forty-two hundred characters known or thought on strong evidence to be inherited in simple Mendelian patterns.[7] Most of those are, like alkaptonuria, metabolic blocks. We know, as Garrod surmised, that these are caused by failure to produce a specific enzyme--that is, the protein molecule that acts as catalyst of the blocked biochemical step.

In 1903, Walter Sutton in the United States and Theodor Boveri in Germany independently made the connection between the patterns by which pairs of Mendel's elements assort themselves and the precisely similar sorting and recombination of the chromosomes in formation of the germ cells and fertilization of the egg. Mendelian elements travel like chromosomes; they must be on the chromosomes. But Sutton also pointed out that the chromosomes in a cell are far fewer than the number of distinct characters to be accounted for. (Fruit flies have four pairs of chromosomes, humans 23.) Undoubtedly, many elements must travel and be inherited together, not sorting freely as do Mendel's seven pairs of traits in peas.[8]

Then between 1905 and 1908 genetic analyses by William Bateson and his colleague R. C. Punnett, among others, showed that some genes specifically modify the action of other genes. Working with the colors of cocks' combs and of sweet-pea blossoms, they demonstrated that more than one gene may be required to produce a normal character. They also observed that in their crosses some characters seemed to be inherited together, yet not with perfect fidelity; they could make no sense of that.[9] In 1910, Edward M. East pointed out that inheritance of a character that appears to vary continuously--in seeming opposition to the round-or-wrinkled, this-or-that, nonblending Mendelian pattern--can be interpreted as Mendelian if one supposes that it is controlled by several pairs of genes whose effects can add up.[10] East demonstrated this with the number of rows, which is variable, on ears of corn, but his conclusion applies to such common phenomena as height (in many organisms other than peas), skin color, and so on.

Mendel's model had not been overturned, yet was distinctly altered: genes are on chromosomes, and may be linked in hereditary transmission or related in biochemical expression. But no particular gene had been associated with any particular chromosome. In 1910, Thomas Hunt Morgan, at Columbia University, was breeding fruit flies, Drososophila melanogaster, which from a single mating will produce, in less than three weeks, hundreds of offspring themselves ready to breed. Morgan had begun in science as an embryologist, concerned, in other words, with how the fertilized ovum, one cell, develops into the juvenile and then the adult organism with perhaps billions of cells that are differentiated into many scores of markedly specialized cell types--muscle, nerve, retinal, blood, gland, and so on. But the problems of embryology seemed intractable, and after Mendel's rules were rediscovered Morgan turned to the transmission of those characters whose expression was so baffling. He was interested, among other things, in the question of what determines whether an individual is male or female. Gender, when you think about it, behaves much like a Mendelian character-pair. Cytologists knew, also, that higher organisms have a pair of chromosomes that is anomalous, in that the two don't look alike, one of them being conspicuously rudimentary. One sex (in flies and humans, the female) has two of the normal chromosomes of that pair, designated X and X; the other instead pairs the normal with the rudimentary chromosome, X with Y. In a line of Morgan's flies that had bred pure for twenty-odd generations, he noticed a male fly with white eyes; normal flies have red eyes. He tried the standard Mendelian crosses, and found that the gene for white eye was recessive to red--and, in an unheard-of result, was inseparable from the factor for sex. That is, it would show up in a male only when inherited from a female (and she could have red eyes if hybrid); yet a female could be white-eyed if her father was (and if her mother was at least hybrid).[11] In brief, the gene for white eye or its alternative, that for red, travelled with the X chromosome, the sex determinant. We now say it was sex linked or, more clearly, X linked.

Morgan's discovery of X linkage was soon seen to straighten out the bookkeeping for the inheritance of other characters. The first example found in humans is red-green color blindness.[12] Another standard example is hemophilia, which entered most of the royal families of Europe by way of Queen Victoria and her many daughters.

Hundreds of human characters are now known to be X linked, including, among the most recently isolated genes, those for Duchenne muscular dystrophy, for chronic granulatomous disease, which is a blinding disorder of the eyes, and for one form of inherited tendency to manic depression.[13]

The gene for white eye in fruit flies was the first to be assigned to a specific chromosome. It led directly to the first genetic map--which was also the next and most important improvement in the Mendelian model. Within a year or so, Morgan and his colleagues had identified a number of X-linked characters in flies. They promptly did the obvious, carrying out matings to see whether these characters are inherited in sets, as predicted from the idea that they reside on the same chromosome. The result of the matings was that linked genes do indeed travel together--but not always. For example, Morgan crossed a white-eyed fly that had normal wings to a red-eyed fly with rudimentary wings, a new X-linked character. After two generations, he was able to get a certain percentage of flies that had white eyes and rudimentary wings: the two mutant characters were now in the same flies, yet still showed the tell-tale X-linkage pattern of inheritance by males from female parents only.

The explanation of this result had to be that an interchange of material sometimes takes place between two chromosomes of a pair. The interchange was termed crossing-over. Imagine a pair of chromosomes as two strands of cooked spaghetti, one tomato-pink the other spinach-green: if the two were to tangle, break at a point of overlap, and rejoin, this would yield one strand that was long pink then short green and the other short pink then long green. That's crossing-over, except that the occurrences were not then detectable by inspection.

In the case of the white-eyed fly, of course, crossing-over could take place only in the female, since the male has but one X. But the idea of linkage that is sometimes broken by the interchange of genetic material obviously explained other anomalous associations of traits not necessarily on the sex chromosomes, going back to Bateson and Punnett's problematic results with sweet-pea blossoms.

In 1913, Alfred H. Sturtevant, a student of Morgan's at Columbia, examined the tables of mating results for fruit flies with six different mutant factors each known to be recessive and X linked. He traced the transmission of each mutation and its normal alternate in relation to each of the other mutants, and thus calculated the exact percentage of crossing-over between the genes taken two by two. Some factors were more tightly associated than others. If crossing-over was indeed due to chromosomes exchanging portions of themselves, then two genes lying adjacent to each other would have less likelihood of crossing over than two lying far apart. Sturtevant realized that the full table of crossover frequencies reveals the order in which the genes are arranged on the chromosome; but more than just the order, it reveals their approximate relative distances from one another. These points established, the tables of data could be extended to take the genes three at a time. This showed that sometimes two genes lying far apart remain linked while a gene in between exchanges with the other chromosome: crossing-over sometimes happens twice to the same chromosome pair. In effect, a double crossover interchanges segments from somewhere in the middle of the chromosomes. Sturtevant also found triple crossovers. All these, however, confirmed the sequence of the genes and the assignment of their relative positions. His paper, "The Linear Arrangement of Six Sex-linked Factors in Drosophila, as Shown by Their Mode of Association," includes tables of data, charts of crossover frequencies, and a single illustration--one long horizontal line with three genes shown near each other at the left end, two genes located slightly to the right of center and quite close to each other, and the sixth at the extreme right. Sturtevant had drawn the first gene map.[14]

The map of defects

Geneticists soon realized that mutations of any particular gene occur rarely and at random; a mutation when it does occur is likely to be recessive to the trait the organism normally exhibits in the wild. Mutations of X-linked characters are peculiarly available for discovery and mapping because a recessive gene on the male's one X chromosome can't be masked by a dominant. Thus, the geneticist can spot rare mutations promptly, can assign them unambiguously to the correct chromosome, and can track them more easily in relation to other genes. None the less, Morgan's team soon found characters not X-linked that associate in heredity, constituting linkage groups. Other geneticists found similar patterns in other organisms. To define an organism's linkage groups is to solve a puzzle: as characters are added to one or another, say A C B + E or L J G K + H, some groups turn out to overlap in several members, say D and I, and coalesce. In such fashion, sure enough, the genes of fruit flies proved to fall into three linkage groups besides the X-linked set. Characters that lie in different linkage groups assort freely, just as in Mendel's peas, while the genes or their mutations within any one linkage group are indeed more or less tightly associated, crossing over with various frequencies. These discoveries of course generated maps--straight lines bristling like barbed wire with the marks showing relative positions of genes along them.[15]

A decade after the first gene map, Sturtevant came upon a rare instance where crossing-over does not hit at exactly equivalent points on the two chromosomes, one winding up, instead, with extra genetic material and the other with a deficiency. In such a case, in the next generation, one offspring may have a chromosome with two copies of some gene, while a sib may lack that gene. He then found that "two genes lying in the same chromosome are more effective on development than the same two genes when they lie in different chromosomes"; in other words, the doubled gene produces an exaggerated character. He called this "position effect."[16] Other studies of linkage and crossover frequencies turned up data showing that groups of genes are sometimes rearranged and even inverted on the chromosome.

From the start, the genetic map of any species, molds or flies to maize or humans, has been primarily the map of defects. Even in the shelter of a milk bottle with plentiful food in Morgan's fly room at Columbia University, those white-eyed flies reproduced less prolifically than their normal brood-mates, while flies with gross mutations of, say, wings, mouth, or thorax obviously could never survive in the wild. In all creatures, almost all mutant characters are deleterious to some degree; indeed, many are lethal. Yet only a mutation--that is, an alternative to the wild type--could identify a unit character and open its heredity to analysis and its gene to mapping. (Victor McKusick, who is the premier geneticist of man, recently called this aspect of the human genetic map its "morbid anatomy."[17])

Two questions arose inescapably: What causes mutations? and What exactly does a mutation change? Morgan himself resolutely refused to consider the nature of the substance of the genes--their chemistry or physics. His most original student was Hermann J. Muller; in 1921, before an audience of naturalists in Toronto, Muller read a paper about the nature of the gene--a paper of astonishing prescience. He conceived of the gene as a particle that despite its ultramicroscopic size must exhibit a complex structure of different parts. He drew geneticists' attention to the dual functions of the gene--its power to direct the organism to make a specific substance, be it enzyme, antibody, pigment, or whatever, that is different from the gene itself, and its power to dictate the formation of an exact duplicate of itself. That power of self-replication is an essential of the gene and was more remarkable than biologists realized, he said--for when a character mutates, the changed gene, too, is henceforth duplicated faithfully. In this fact, years ahead of most other biologists, he saw the clue to Darwin's predicament, the fulfillment of Mendel's promise to evolutionary theory:

> Inheritance by itself leads to no change, and variation leads to no permanent change, unless the variations themselves are heritable. Thus it is not inheritance and variation which bring about evolution, but the inheritance of variation, and this is due to the general principle of gene construction which causes the persistence of autocatalysis despite the alteration in structure of the gene itself.

Given that sort of structure, Muller said, "evolution would automatically follow."

Understanding the gene as a substance with structure, Muller distinguished possible types of mutation in it--loss, gain, substitution, or rearrangement--and noted that other types of mutations, changes of the genetic material on a larger scale, must also take place. He said, "Some new and unusual method may at any time be found of directly producing mutations." He expressed simultaneously a commanding view of the central position of genetics in biology and the practical necessity to determine the physical nature of the gene--"to grind genes in a mortar and to cook them in a beaker, after all."[18] Then in 1927, Muller demonstrated that X rays cause mutations in fruit flies. The frequency of mutations increases with the dose of X radiation; but which genes will be affected is not predictable.[19] The next year, Lewis Stadler showed that ultraviolet radiation can cause mutations. (By the 1940s, mutations were also known to be caused by a variety of chemicals, in a catalogue that still grows longer by the week.)

In 1931, two American geneticists, Barbara McClintock and her student Harriet Creighton, working with the genetics of maize, published a curious observation. They had constructed a strain of corn in which the second-smallest chromosome had an irregular shape--a knob--visible under the microscope. In other words, a chromosome other than one of the X-Y pair could be distinguished from its normal pair. They were able to identify genes for three characters that were associated with the knob, by demonstrating that when crossing-over took place that separated these genes, the knobbed region showed up in the chromosome sets of those descendents carrying the associated traits. The genes and that part of the chromosome stayed together.[20] The result nailed down the fact that genes are arranged on chromosomes. More generally, it showed that the genetic material may have distinguishing anatomical features.

In 1933, in a brief paper in Science, T. S. Painter announced that he had charted perceptible differences among chromosomes under the microscope--differences detailed enough to correlate crossing-over of genes as shown in the statistical tables with physical interchanges in the material of the chromosomes. The larvae of many species of flies, including Drosophila, have salivary glands that contain abnormally large chromosomes, elongated and segmented in a series of rings, or bands. When such a chromosome is stained with the appropriate dye, Painter found, the bands display a fine sequence--light and dark, thick or thin, bulging or pinched. His drawing of one chromosome looks like an assortment of washers stacked on a knitting needle. He saw that the banding gives each pair of chromosomes a characteristic pattern, the same from fly to fly, differing from one pair to the next. For the first time, chromosome pairs other than the X and Y could be distinguished unambiguously. More, rearrangements of portions of any chromosome--single or multiple crossovers, doublings or deletions, translocations or inversions--could be detected and correlated with genetic changes.[21] In a burst of papers in the next months, Painter and his students assigned the Drosophila linkage groups to chromosomes and located many individual genes on the chromosomes with precision. To be sure, the method appeared to be restricted to Drosophila and related species. Other organisms offer no giant chromosomes.

The map and the pathway: genes and enzymes: genetics and biochemistry

Essential to understanding the expression of genes, from Garrod's conception of a genetically determined metabolic block and onwards, is the place of the gene in the biochemical processes of the organism. Biochemistry was the other of the biological sciences that developed in the decades around the turn of the century. The central concern of biochemistry then and now has been the functioning of certain of the large molecules active in the chemistry of the organism, and preeminently the enzymes. The living cell contains an immense variety of different molecules, but for the most part they are either quite small or extremely large. The large molecules--as biochemists deduced by the 1890s and established beyond doubt by the end of the 1930s--are long strands built up of subunits linked together. Such chains, called polymers, are rare in nature except as the products of living things, where they are common. (We also make artificial polymers, of which one of the earliest, of course, was nylon.)

The large biological molecules are of four kinds: lipids, meaning fats; carbohydrates, meaning chains of sugar molecules; proteins; and nucleic acids. Lipids and carbohydrates, however long their chains, are chemically and conceptually boring--monotonous repetitions of one or a few subunits in entirely predictable order. The nucleic acids are deoxyribonucleic acid, now familiar to schoolchildren as DNA, and ribonucleic acid, or RNA. They were discovered late in the nineteenth century. They are composed of four kinds of subunits, called nucleotides. Each nucleotide has projecting from it a flat chemical entity called a base. The bases in DNA are of four sorts, differing slightly: adenine and thymine, guanine and cytosine, usually symbolized A, T, G, or C. Until the late 1940s everyone supposed, though the evidence was meagre, that the four nucleotides were repeated over and over again, in a structural unit called a tetranucleotide, making nucleic acids much like carbohydrates. Max

Delbrück, one of the founders of molecular biology, observed that nucleic acids had been taken to be "stupid molecules"--substances that could perform no interesting function, molecules that could carry no information.[22]

The interesting molecules were the proteins. These were known to do virtually everything in the biochemistry of the organism. Many of the hormones are proteins. Immunoglobulins, the antibodies that are an essential part of the system that protects us against foreign substances including viruses, bacteria, and some cancer cells, are proteins--as, indeed, are most of the substances they react against. Hemoglobin is a protein, as are a whole set of other respiratory molecules in animals and plants. But the quintessential function of proteins is to act as enzymes. The cell breaks down or builds up its component parts in many series of chemical reactions, step by step, in a network of metabolic pathways of stupendous intricacy. Each step in every pathway is made to occur quickly--thousands of times per second per cell for some reactions--by the action of an enzyme, a catalyst, specific to that step. Francis Crick pointed out to his fellow biologists thirty years ago, "It is at first sight paradoxical that it is probably easier for an organism to produce a new protein than to produce a new small molecule, since to produce a new small molecule one or more new proteins will be required in any case to catalyse the reactions."[23]

In first approximation, then, the problem of the genetics of gene expression was to determine the relationship of genes to enzymes and other proteins.

One gene, one enzyme

Garrod's initial understanding of inborn errors of metabolism prefigured a line of research into the expression of genes that was pursued, for some time, in partial independence of the transmission genetics of the Morgan school. Muller knew Garrod's work; many other American geneticists then did not. A contributory reason was that Garrod and many of his successors were English, beginning with Bateson and Punnett and their demonstration that several genes are necessary for expression of color in blossoms. Yet perhaps the most influential early worker in this line was an American, Sewall Wright, who trained at Harvard, in the period just before the first world war, where a group of geneticists were working not with flies but with small mammals. In a series of papers in 1917 and 1918, Wright analyzed the inheritance of coat colors in guinea pigs, mice, rats, rabbits, horses, and other mammals. He stated his program:

> It remains for genetics to assist embryology and biochemistry in filling in the links in the chain between germ cell and adult in specific cases. Variations of adult characters must be traced back through the contributing causes at each stage of development . . . and, on the other hand, the ramifying influences of unit variations in the germ cell must be traced forward through development.

In particular, Wright showed that production of the pigment making for coat color in mammals requires that several biochemical steps take place in sequence; he surmised that each step is mediated by a different, specific enzyme. By drawing attention to the specific enzyme related to the gene, and to the sequence in which enzymes act, he established the subscience of biochemical genetics.[24]

In doing so, Wright set up several presumptions. The first is a point suggested in earlier work which he now stated with power and generality: that genes produce characters in combined action, meaning that genes can have secondary influences and that characters typically are modified by actions of more than a single gene. (We now say that in the development of the organisms many genes are pleiotropic and many characters polygenic.) The second presumption is that the gene is expressed through the action of an enzyme. The third, implicitly, is that some relationship ought to obtain between the map of genes on the chromosome and the sequence of steps in the metabolic pathway. These were necessary approximations in the continuing development of the Mendelian model, and demonstrate the complexity and importance of the idea of genetic maps and sequences. Wright continued in this research---in a span of interests that included population genetics and evolutionary theory--for forty years and more.[25]

In the summer of 1935, in France, a young American from Morgan's lab named George Beadle collaborated with the leading French geneticist of that period, Boris Ephrussi, in a set of experiments using two microscopes and four hands in which they transplanted tissue between the larvae of fruit flies bred with various mutations of eye color, and then watched the mature flies develop. They thus reversed biochemically the effects of the particular genetic crosses, and determined the sequence of successive steps, catalyzed by whatever was produced by two specific genes, in the metabolic pathway that leads to the normal pigment of the wild-type red eye. They wondered whether the product of each gene was some hormone. Back in the United States, Beadle began to seek out better organisms with which to do genetics--simpler, more easily manipulated creatures with a faster generation-time than flies. He and a colleague, Edward Tatum, settled on a one-celled organism, Neurospora, which is a red mold that commonly grows on bread in the tropics. They found mutant strains of the mold, deficient in one or another factor necessary for growth, and were then able to carry out in a day or two experiments in biochemical genetics that with flies would take months. By 1941, their work appeared to have convinced most geneticists of the idea that had long been in the air, that what the gene does is to specify an enzyme. This they called the one-gene-one-enzyme hypothesis.[26]

Biochemistry and the sequencing of biological polymers

A fundamental transformation took place in the late 1940s in the understanding of the chemical nature of large molecules of biological importance. Through the 1930s and virtually until the end of the forties, biochemists took it for granted that genes must be made of protein. Hormones, immunoglobulins, enzymes--proteins and only proteins were known to be capable of the specificity that genes display. But then in the mid-forties the possibility emerged that the material of the genes was not protein but a different class of macromolecule altogether--the nucleic acids, in particular DNA. Yet for some time this new identification of the genetic material explained little: proteins and nucleic acids alike were poorly characterized, their three-dimensional structures unknown, the chemical basis for biological specificity unguessed.

Despite that specificity, when biochemists of that era turned to the substances of the molecules, to their chemical makeup, they thought of them in general, nonspecific ways. Nucleic acids were thought to be long strings of the four different nucleotides repeated over and over. Proteins were known to be chains built up of assortments of a score of different kinds of amino acids--yet they were similarly believed to be multiples of fundamental units or spacings many times repeated. The amino acids were supposed to be strung together according to some general chemical rule--complex, no doubt, but predictable. For proteins and nucleic acids alike, these ideas represented the traditional direction in which explanations of the chemical structures of these substances were to be sought.[27]

The work of two biochemists proved decisive in changing the way people thought about specificity: Frederick Sanger and Erwin Chargaff. In the course of nearly a decade, working at the University of Cambridge beginning in the mid-forties, Sanger used a new set of techniques called chromatography to determine the amino-acid sequences of the bovine insulin molecule. By chromatography and related methods, vanishingly small amounts of virtually identical complex substances can be persuaded to separate themselves for identification and measurement as they migrate, in solution, along a sheet of filter paper, down a column of inert granules, or across a slab of hard gelatine, some travelling faster and farther than others because of differences in weight or solubility or electrical charge among molecules. When Sanger began with insulin, not even its molecular weight was correctly known. The molecule was thought to be twice its true size and composed of four chains of amino acids; the composition of the chains was taken to be heterogeneous.

Sanger showed that the molecule is of two chains, linked by cross bonds at certain locations, and that the sequence of each chain is unique and always the same, meaning that every molecule of insulin in a given cow is exactly like every other. These results came slowly. He first worked with fairly rough chemical methods for knocking off and identifying the amino acids at the tips of the chains. He then realized that natural digestive enzymes will break a protein chain at specific places, each type of enzyme cleaving only the bonds between two particular amino acids, and nowhere else. With a toolbox of such enzymes and some radioactive labels, he was able to break the molecules into several overlapping sets of fragments, separate the fragments chromatographically (and find each group on the filter paper by its radioactive label), then analyze each fragment into its components. Finally he related the fragments to each other by their overlaps as though solving a word puzzle, to get the sequences of the chains and the entire molecule.

Even as Sanger worked, the news slowly spread and its implications sank in. The conclusion was forceful by mid-1949, when he went to the annual symposium on quantitative biology at Cold Spring Harbor, Long Island, summertime headquarters of Delbrück and his associates. In a paper published on the first of June of that year, he was already able to say that "there appears to be no principle that defines the nature of the [amino-acid] residue occupying the particular position in a protein."[28]

Which is to say, the sequence was entirely and uniquely specified. No general law, no physical or chemical rule, governed their assembly. Sequence specificity in the protein demanded a specific instruction from the gene.

In 1950, Erwin Chargaff, who five years earlier had begun using chromatographic methods to analyze nucleic acids, published a paper reviewing his results. He and his colleagues had analyzed the DNAs of a variety of different organisms. The four kinds of nucleotides appeared in DNA in proportions that were constant in a given species but that varied widely from one species to another. Only in a few species did they approach equal representation of the four. Chargaff's paper demolished the notion that DNA was a repetitive molecule, a "stupid substance" that could not carry specificity. He established that DNA, after all, could be as specific in sequence as proteins.[29]

Molecular biology and the structure of the gene

The first concern of molecular biology has always been to understand the nature and action of genes and their products, in terms of their physical structures as assemblages of atoms. One may define molecular biology as the pursuit of the structural basis for the specificity of genes and their products. From the early 1940s, scientists of that small band who later called themselves molecular biologists began by attempting to analyze the transmission genetics of the simplest creatures they could find. Where Beadle and Tatum were using *Neurospora*, the group around Delbrück worked with a strain of the common intestinal bacterium *Escherichia coli* and a set of the viruses, called bacteriophage, that prey upon *E. coli*. Bacteria are one-celled creatures without cell nuclei. Together with certain algae, they are called *prokaryotes* (from Greek roots meaning "pre" and "kernel" or "nucleus"). The total of different biochemical steps E. coli carries out--the number of different enzymes, hence of genes--is about two thousand, three or four per cent of the number performed by the person in whose gut the bacterium resides. Bacteria can reproduce, by doubling, every half hour or less, achieving in a day as many generations as fruit flies produce in several years, more than man has known since the Crusades. Viruses are still smaller and simpler, with a dozen genes or less. They are not able to reproduce by themselves. Instead, a virus invades a host cell, captures that cell's machinery and turns it to reproducing more virus. Bacteriophage can multiply a hundred-fold in twenty minutes. Nor do genetic experments with bacteria and viruses typically require a microscope. These rates of doubling mean that when one or a few bacteria are seeded onto a petri dish--a flat, low-sided glass plate--of nutrient jelly, within a few hours the colony grows to a spot visible to the naked eye; when a virus is introduced to the culture its progeny soon betray their presence as they clear a disc of burst cells in the bacterial lawn. Molecular genetics has been low in budget, high in ingenuity.

In the early 1940s, though, bacteriologists generally supposed that bacteria had no genes: they were not thought part of Mendel's realm. One of Delbrück's first discoveries, made with Salvador Luria, was the demonstration that bacteria mutate: their genetic nature was proved. At about the same time, independently in German-occupied Paris, a French microbiologist (and Resistance leader) named Jacques Monod made the same discovery by a different method.[30]

The three-dimensional structures of nucleic acids and proteins, on the other hand, had to be approached through a technology developed in the physical sciences--X-ray crystallography. The method was invented in 1912, when it yielded the arrangement of the atoms of sodium and chlorine in crystals of common salt. Physical chemists used it through the 1920s and 1930s to get out the precise spatial arrangements of the atoms composing molecules of ever-greater size and complexity, so long as those molecules would form orderly crystals. By the end of the 1930s, when war interrupted, several laboratories had turned the methods of X-ray crystallography on crystalline forms of proteins and of nucleic acids. In 1951, James Watson, an American post-doctoral biologist trained in the Delbrück group, came to England, to the Cavendish Laboratory, the experimental physics laboratory in the University of Cambridge, where he joined a small unit that was trying to determine the structures of protein molecules. There he met Francis Crick, a physicist beginning to work on biological questions. (Crick had been acutely aware of the significance of Sanger's earliest results on the amino-acid sequence of insulin.) Eighteen months later, in the spring of 1953, Watson and Crick built and published a model of deoxyribonucleic acid, DNA--since celebrated as the double helix.[31]

The double helix of DNA is a conception of astonishing parsimony in the relation of structural detail to functional necessity. It brings together the two lines of genetic inquiry, transmission and expression. The strands of DNA twine coaxially, clockwise, the one running up, the other down. Each strand is a linear sequence of nucleotides--knuckles forming the backbone of the strand, each with its base--adenine or thymine, guanine or cytosine. The two strands are at the outside of the structure, and are hooked together across the space between them by the bases, which form pairs. The base pairs are thus the ribs of the structure. The bases are all-important, because their sequence along the strand is the only variable part of the structure. The crucial fact that Watson discovered is that the physical shapes of the bases limit them to just two kinds of pairs as they connect the strands: they pair with almost perfect fidelity A to T, and G to C. The bases are spaced along each strand so that exactly ten pairs occur in the length of a full turn of the helix. The pairs are stacked flat, with

3.4 angstrom units and a tenth of a revolution separating a pair from the one above or below. Three-point-four angstroms is about one-and-a-third hundred-millionths of an inch.

In one dimension, the freedom afforded by the structure is almost unimaginably generous: along a strand, the bases can, in principle, appear in any sequence--and though they are but 3.4 angstroms apart, while the diameter of the double helix is twenty angstroms, the total DNA in a typical mammalian cell measures more than a metre. In the other direction, the constraints are almost perfectly stringent: between the two strands, once the sequence on one is fixed, the base pairing--adenine to thymine, guanine to cytosine--altogether determines the complementary sequence on the opposite strand.

Form and function are one. The model illuminated at a glance the mystery of heredity and the mystery of growth and development--the passing on of potential and its expression. The constraint imposed by the pairing rules makes it a unique consequence that if the strands part, each can assemble on itself a replica of its complementary previous partner, producing two identical double helices at the time of cell division. That's gene transmission--in principle, at any rate, for it took several years to show that in life DNA does replicate this way and several decades to work out the biochemical details. The freedom of the sequence along the strand allows the encipherment, as with a four-letter alphabet--A, T, C, G--of the entire specification for the substance and regulation of the organism. That's how the gene dictates the organism--again, in principle. Proteins, though different in chemical makeup, are also long strands of building blocks; the catch was that though the bases are of four kinds, there are twenty different essential amino acids. Thus, what the gene, written in a four-letter alphabet, must do is determine a corresponding sequence of the twenty amino acids. That was the problem stated simply and abstractly. It was soon called the coding problem. The other face of the problem, messy, biochemical, was to elucidate the mechanics of the translation from DNA into assembled protein.

Sexual recombination, coitus interruptus, and the map of bacterial genes

Most nonspecialists have only a partial view of how classical molecular biology achieved its results. In particular, laymen and historians alike tend to misunderstand the importance of the discovery of the structure of DNA. The structure has elegant explanatory force. Its discovery was a marvellous drama. Of course DNA is dazzling. But in the development of the science another discovery, made during the several preceding years, was equally instrumental. This was the discovery of bacterial sex.

Even when the genetic nature of bacteria (and viruses) was proved, by the demonstrations in the mid-1940s that they mutate, bacteria were known to reproduce asexually, the cell splitting into two identical offspring. But then in 1946, Edward Tatum and Joshua Lederberg showed that bacteria sometimes do exchange genetic material directly, in a process they named conjugation. By 1952, William Hayes, in London, showed that what happens in conjugation is that one cell (promptly termed the male) pipes a copy of its genes into the other (the female). This transfer takes some ninety minutes. Meanwhile, in 1951, Lederberg and Norton Zinder showed that bacteria sometimes exchange genes by another, indirect method, which they termed transduction; here, a virus mediates, by snaring bits of DNA from one bacterial cell and transporting the bacterial genes into the next cell it infects. Then in 1955 Elie Wollman and François Jacob, at the Institut Pasteur in Paris, showed that bacterial conjugation could be stopped at any moment when only some part of the genes had been piped across. They did this by agitating the culture violently in a Waring blender, and called the experiment, of course, "coitus interruptus."[32]

These practices provided the bacterial equivalent of matings in fruit fly, maize, mouse, or man. Indeed, they were more powerful and precise than matings in higher organisms, for they enabled geneticists to move small, selected portions of genetic material into a bacterium and then observe the outcome--seeing just how the inserted genes expressed themselves and how they were switched on and off. These methods have led to most of the major discoveries of molecular biology. In particular, geneticists employed them to map the genes of microorganisms at unprecedented high resolution. More generally, the methods turned out to be adaptable to the developmental or biochemical genetics of higher organisms.

The idea of a map or sequence of genes was rapidly becoming more precise yet more complex. At the Institut Pasteur, Jacob and Wollman were able to clock the transfer of particular genes from donors of one strain of E. coli into recipients of another by breaking off conjugation after various lengths of time and then observing the recipients for various biochemical mutations they had not possessed. (The work was simple in principle, tricky in practice; for example, they had to arrange deficiencies like sensitivity to antibiotics that would inactivate the males in the culture after conjugation was interrupted.) They found that in any given strain the map of genes was always the same, and they displayed the map as a dial marked off in ninety minutes.[33] Over the years, the dial has become crowded, with some segments spacing genes only seconds apart. Yet in principle the map of E. coli's genes is a direct descendant of the first maps of Drosophila's X chromosome, for it depends on breakages and recombinations in the chromosome, detectable through comparing the action of gene mutations with the wild type.

Meantime, the discovery of the structure of DNA magnified the gene. It was no longer a point on the chromosome but rather a length of base pairs--and, therefore, it has an internal sequence. Thus, among other things, the structure of DNA also implies a theory of mutation. A gene can mutate in several ways: by deletion or addition of one or more base pairs, or by substitution of one base pair for another. Further, breaks and crossing-over, or other recombination events, should hit not merely between genes but within them. In the mid-1950s, Seymour Benzer, then at Purdue, a physicist recently recruited to the Delbrück group, devised an experimental set-up to map mutations within a short genetic region of a particular bacterial virus. He found a way to infect bacteria with two different virus mutants simultaneously; in the bacteria these would trade segments of DNA in a way exactly analagous to crossing-over in fruit flies. With many hundreds of mutants and many thousands of such matings, Benzer over a five-year period mapped recombinations of genetic material that often distinguished between mutational changes that had taken place at adjacent base pairs.[34] The gene had a fine structure that could be mapped by classical genetic techniques cleverly adapted.

New kinds of genes: structure _vs_ control

In the mid-1940s, Barbara McClintock had begun to see in the pedigrees of her corn plants indications of genetic elements that were not ordinary genes yet that influenced ordinary genes--and not the character expressed but rather the rate of expression. Genes were being turned on or off, amplified or diminished, by other genetic elements. She found inhibitors, inducers, and others, which she soon grouped together as "controlling elements." Some of these, she divined, move unpredictably from place to place on the chromosomes. She began to publish papers about controlling elements in the early 1950s,[35] but they were largely ignored. Many of her ideas appeared impossibly bizarre.

At that time, however, in Paris, François Jacob and Jacques Monod began a collaboration, working on the expression of genes in *E. coli* that enable the bacteria to utilize galactose, a complex sugar, breaking up the molecules for energy and raw materials. Over the next several years they found that the bacteria require three different proteins to digest galactose. But the curious fact was that the bacteria don't make the necessary enzymes unless galactose happens to be available. This brought them to a general question.

Even a bacterium, and still more a cell of a higher organism, has the genes for many more functions than it carries out at any given time. Indeed, if the cell did do all those things all the time it would rapidly exhaust the energy supply. But how does the cell turn functions on or off just as needed? Jacob and Monod found that *E. coli*'s three genes for galactose metabolism lie next to each other in a fixed sequence on the bacterial chromosome, and that certain other elements are right next door. The first of these is a master turn-on for the whole series of genes, which they called the *operator*. They found other control functions as well. And they located a gene for still another function, not at all an enzyme but a protein that actually sits on the DNA at a specific place. This molecule locks the sequence of genes off. But when a galactose molecule appears, it plugs into another spot on the protein, acting as a key: the lock drops off the DNA and the sequence of genes can be read out. When all the galactose is digested, even the key, the protein molecule re-attaches to the DNA and turns the sequence off again. Jacob and Monod christened the entire coordinated sequence of genes and controls on the chromosome the *operon*. They called that controlling protein the *repressor*.[36]

These discoveries were of highest importance. They established the existence of genetic regulation--of control functions located on the chromosome, mappable, in the

DNA sequence. Henceforth, genes that read out as enzymes or other proteins have been distinguished, where necessary, as structural genes, as opposed to the new-found regulatory genes and other control elements. Further, they established the existence of proteins that have dual specificities, responding to the presence or absence of one chemical substance (say, galactose) by reacting with another, chemically unrelated substance (say, a binding site on the bacterial DNA). Monod and Jacob recognized other repressors, controlling other functions in bacteria and viruses. They saw that some enzymes, and even carrier molecules like hemoglobin, have similar complex specificity; indeed, such proteins have many functions in the organism. Still further, regulation of metabolic pathways in response to varying conditions requires biochemical versions of what information theorists and computer engineers had begun to call feedback loops--that is, systems whose outcome or end product influences the continuing earlier stages of the process. Monod thought that that particular, negative feedback model provided by the galactose operon and repressor was so elegant it would prove universal, but the feedback loops in living systems have turned out to be of every imaginable kind.

The necessity and ubiquity of regulatory genes and other control loci as part of the very genetic sequence creates great new complications for the mapping of genes.

The meaning of genetic sequences: the messenger and the genetic code

The work in Paris on the regulation of the expression of genes also threw off essential clues to the molecular machinery by which the cell constructs the protein dictated by the gene--that is, clues to the exact process by which the sequence of base pairs on the DNA is converted into the correct sequence of amino acids in the protein chain. In the 1950s, electron microscopy had shown the insides of cells to be filled with anatomical structures, including an enormous quantity of a complex molecular organ soon termed the ribosome. Biochemists then demonstrated that ribosomes are the sites where proteins are assembled; they are tiny knitting machines that select amino acids from the cellular soup and link them together. Jacob, Crick, and others put the final missing component of this mechanism into place when they realized that the rates at which genes are turned on and off show that an intermediary must carry the genetic instructions from the DNA to the ribosomes. This intermediary is, in fact, a length of DNA's chemical cousin, ribonucleic acid. RNA is also a string of bases, but in single strands. In different forms it carries out a variety of functions in the synthesis of proteins by cells. To read out the gene, a new RNA molecule assembles along the DNA, the RNA bases lining up by the complementary base-pairing rules. This step is now called transcription, and the assembled long strand is called messenger RNA, or mRNA for short. The messenger is then picked up at the correct end by a ribosome, which acts as a reading head, accepting its instructions for the sequence of amino acids. This step is called translation. The steps in the process were clear by the summer of 1959, and were published the next year.[37]

The specificity of genes thus turned out to be radically different from the specifity of proteins. In principle, their mystery was now solved. The specificity of the gene is informational, linear, one-dimensional. A structural gene is an instruction, a message, and it is written as a unique sequence of bases on the DNA; it specifies a unique sequence of amino acids in a protein chain. In contrast, the specificity of the protein molecule is conformational, three-dimensional. Be the protein an enzyme, a hormone, a repressor or other regulatory molecule, hemoglobin, a cell-surface receptor, or an immunoglobulin, its specificity results from the contours, pockets, and local distribution of charges over the surface, which result when the completed chains of amino acids fold up into their precise final shape. But information dictates conformation. The one-dimensional sequence dictates the three-dimensional form. The protein chain, once correctly strung together, folds itself automatically--by virtue of the shapes and charges of that particular sequence of subunits--into its correct specific functioning shape.

The coding problem remained: What particular combinations of the four bases in fact specifies each particular amino acid? The summer of 1960 saw the first break in the coding problem. Two young scientists at the National Institutes of Health made an artificial strand of RNA in which all the bases were of just one kind. They put that as a messenger into a system complete with ribosomes, amino acids in solution, and other cellular components necessary for protein formation. They got out a protein chain in which all the amino acids were of just one kind. Over the next six years, artificial messenger sequences proved the key to the full dictionary of the genetic code.[38]

A triplet of bases in a particular order along the nucleic-acid strand specifies an amino acid. Such a triplet is called a codon. The bases cytosine-cytosine-guanine (CCG) on the messenger, for example, cause the ribosome to add the amino acid proline at the next position in the growing protein chain, while the same bases in reverse order, the codon GCC, translate as alanine. The genetic code has 64 codons (any of four possible bases in the first position, in the second, and in the third) and a lot of redundancy. Three of the 64 codons specify no amino acid but are, instead, stop signals, signifying the end of the chain.

The distinction between structural genes and regulatory functions, and the elucidation of the genetic code, of course redefined the ideas of the sequence and of the map of genes. The sequence becomes simply the linear run of bases, or base pairs, defining codons and making up genes. The map becomes more complex, however--no longer the line of genes on the chromosome but the line of structural genes and control elements, identified and located, the significant points of mutation within each portion also identified and explained, together with an account of the relationships of all these aspects to one another.

The impasse of 1968-1970 and the emergence of recombinant DNA

At the end of the 1960s, the founders of molecular biology were cocky. In the preceding twenty years, they had erected a coherent and satisfying framework of explanation for the genetically controlled processes of the cell. These functioned, they saw, through the dual, complementary structure of DNA; through the information encrypted on the DNA and the mechanisms by which it is read out to direct the building of the organism; and through the feedback loops that control the biochemistry of the cell and the functioning of the genes themselves. The machinery thus revealed was characteristic of all life. At its center was the universal genetic code and those regulatory loops. Structure and function, anatomy and physiology, had been unified at the molecular level.

Or, at least, molecular biologists had achieved all this in outline, for the simplest one-celled organisms, the bacteria and other prokaryotes. An immense amount of detail remained to be filled in, but the founders were confident that the rest of bacterial molecular biology would yield to intelligent drudgery, with no surprises. In this they were correct.

So in 1968 and 1970 molecular biologists were looking for the next fields ready, as they liked to put it, "to go molecular." To begin with, there was of course the necessity to consolidate "the classical stage," as Crick said at the time, by "filling in all of the biochemistry, so that what we know in outline we also know in detail." But then came the new work--and this was of two kinds. One job, Crick said, was to ask "the fundamental classical questions about genes and their products all over again for higher organisms"--eukaryotes, creatures whose cells have nuclei. This would no doubt entail hard work. But the revolution was also to move beyond: it would next strike neurobiology, and development and differentiation (which is to say, embryology), and muscle action, and--some added--the molecular mechanisms that take place at the outer membrane of the cell.[39]

This confident agenda contained a fatal flaw in logic. The problem regarded as tough but routine, doing classical molecular biology all over again for higher organisms, is identical with one of the ripe new subjects, development and differentiation. The molecular biology that distinguishes eukaryotes from bacteria is, precisely and entirely, the molecular biology by which a single fertilized ovum becomes the multi-celled creature of differentiated tissues. This has always been the most intractable problem in biology. It had been intractable at the turn of the century, when modern genetics began. In sixty-five years, that line of attack, developing through biochemical genetics and molecular genetics, had come full circuit and was now facing, once more, the problems of the expression of the genetic information in the growth of eukaryotes. The problem looked as intractable as ever.

In response, the younger biologists reinvented sex.

They didn't put it exactly that way. After all, an obvious distinction of higher organisms is that they reproduce sexually and in the process reshuffle their genes. Sexual reproduction recombines DNA. What was there to reinvent?

A couple of the most pedestrian characteristics of science here came into play. First, scientists are helplessly dependent on their technology--on the development of instruments for observation and of methods for manipulation. A reason good new science is hard is that, by definition, it works just beyond the full stretch of its technology. A reason scientists dislike so-called targeted research, where funding organizations specify goals (as in the "war on cancer" of the early 1970s), is that the prerequisite technological breakthroughs cannot be ordered ahead. Second, scientists are more conservative than people seeing only the results can easily credit. Neither in theoretical speculation nor in those techniques at the bench do they often sail beyond sight of the shore. Most of the time they proceed by small transformations of previous ideas and methods--often by trying rather unimaginative analogies.

Baffled by the problem of doing molecular biology all over again for eukaryotes, in 1970 the new generation of scientists turned to what had worked before, to see whether they could adapt the familiar, successful methods of genetic analysis in single-celled creatures. Eukaryotic cells are far larger than bacteria--proportionately as a horse to a bumblebee. They have thousands of times more genes, and tens of thousands of times more DNA. Yet the problem of development and differentiation can be restated as one of finding the controls that switch whole blocks of genes on and off, in the correct order. Since all the cells in the earliest stages of an embryo have exactly the same gene set, the essence of how cells grow different from other cells necessarily lies with these switches and rate controls on the expression of genes. Only by such regulation can one imagine that of two adjacent cells in an early embryo one is directed to multiply many times to become a liver, while the other multiplies and differentiates further to become a brain.

As a prerequisite, the new molecular biologists taught themselves to grow animal cells not as coherent tissues but as bulk cultures of independent cells, like bacteria. They tried using the viruses that infect animal cells to bring out and transport bits of genetic material, the way bacterial viruses do with bacteria. Especially interesting here were animal-tumor viruses, because they do not kill the cell but take over and redirect its biochemical machinery. The scientists learned to their delight that they could keep right on using bacteria, putting segments of DNA from eukaryotes into them. The bacteria can then be made to multiply normally, doubling and redoubling several times an hour, with the foreign DNA multiplying in step. Thus, starting with a few copies of a segment of genetic material from a higher organism, a biologist can grow bulk quantities of those genes. Molecular biologists took over from botanists the term *clone*, meaning the multiple identical descendants of a single cell, to describe such a culture of bacteria grown up to multiply transplanted genes. By still other tricks, the necessary control functions can be moved with the foreign DNA, or the bacteria's own control mechanisms can be taken over, so that the cloned genes can be got to function in the bacteria--transcribed, translated, and making the protein end-product. Yet another technique for finding genes and genetic functions has been to fuse cells of two different origins--even of two species, such as man and mouse.

Crucial to these methods was the discovery of certain new enzymes that can manipulate nucleic acids. The first of these is the class of enzymes that copy nucleic acids--stringing together nucleotides to form the new complementary strands when the double helix of DNA replicates, or transcribing DNA into messenger RNA. These are called *polymerases*, and cells of various organisms utilize a number of them, DNA polymerases and RNA polymerases. An important small subset of these comprises enzymes that transcribe in the reverse direction, making a DNA copy from an RNA message. Many viruses that infect higher organisms, especially animal-tumor viruses, contain their genetic instructions as a strand of RNA rather than DNA; this they inject into a cell, infecting it. Biologists had thought the RNA was copied directly into RNA as the virus forced the cell's machinery to make more virus. But then in 1969 Howard Temin and David Baltimore, working independently, found that many such viruses, once their RNA is in the host cell, transcribe it into DNA which is then integrated into the host's DNA. They isolated an enzyme that does this, and called it reverse transcriptase. Using such an enzyme, a biologist can start not with a DNA gene but with a strand of messenger RNA, use an RNA virus to put it into a cell, read the message back into DNA, grow up the clone, and then get the gene expressed to see what it is for.

Others, in many different laboratories, in a series of small discoveries of extraordinary cumulative import, found enzymes that break or repair nucleic acids. Cells turn out to have self-protective mechanisms that disable alien DNA, as when a virus invades. These include protein molecules called restriction enzymes (because they restrict the range of hosts in which a virus can survive) which act like submicroscopic wire cutters to snip DNA in precisely specified locations--between particular sequences of bases. Cells have other self-protective enzymes with which they repair their own DNA when it breaks or is miscopied.

In the early years, at least, the aim was always to find ways to move small, identified segments of genetic material from one cell to another, and then to see what difference the inserted genes made. Often, for analytical convenience, the move was from a species that was poorly understood into a different species, typically bacterial, that was well known. Yet rather suddenly--by the end of 1973--molecular biologists realized that they had assembled a remarkable armamentarium.

By such means, for example, at least in principle--the practice is still tricky--large quantities of rare biochemical substances can be grown. Among early successes were production, in bacterial cultures, of human growth hormone and human insulin, both potentially invaluable in medicine and lucrative to their makers. Starting with such methods, also, several teams of researchers are now attempting to find ways to insert particular genes into the cells of infants suffering an inborn error of metabolism, actually to cure the genetic defect. None the less, the technology of recombinant DNA, otherwise called genetic engineering, was invented and is today being elaborated not primarily to make money, nor even to cure disease, but to open up a profound and heretofore intractable problem in pure science. These techniques were invented so that we can find genes in higher organisms and learn what they actually do.

The genetics of eukaryotes is full of surprises. For example, the genome--meaning the entire gene set--of a rat pup or a child is unexpectedly fluid. Genes interact with each other, affecting characters all across the field of development, more intensely and intricately than was ever surmised. McClintock was correct, after all: some genes jump from place to place, even from chromosome to chromosome, and the same gene will have different effects depending on its location. One most unsettling discovery was that in higher organisms many genes are interrupted, in the sequence on the DNA, by stretches of bases that have no relation to the gene, that appear to have no function at all. These are termed intervening sequences, or introns; the stretches of DNA that are part of the structural gene are called exons. The phenomenon complicates the function of messenger RNA. After the gene is transcribed, the mRNA must be edited, so that the introns are snipped out. Only then can the mRNA strand pass through the nuclear membrane to the cytoplasm, where the ribosomes and the rest of the translation machinery are, to be read into protein. The controls on this process of transcription and editing are by no means fully understood. Indeed, control mechanisms in higher organisms are turning out to be unexpectedly different from those in bacteria, and far more diverse. A disappointment was that the elegance of bacterial operons and repressors has no obvious analogue in higher organisms. More surprises are surely in store.

All of these complexities enrich the notion of a genetic map, and emphasize the distinction between a map and a sequence.

The sequencing of biological polymers cont'd: nucleic acids

When Sanger had completed the amino-acid sequences and links of the two strands of the insulin molecule, he turned to the problem of sequencing nucleic acids. The target was obviously important, and many biologists at many laboratories were working at it. Yet it was technically more difficult by a couple of orders of magnitude. To get a significant amount of a protein molecule like insulin or hemoglobin, from a single species of creature and of high purity, is relatively simple. When Sanger began with nucleic acids, though, he had no way to secure a supply of molecules that he could be confident were all alike. Again, the amino acids in proteins are twenty, so that sets of overlapping sequences are easy to distinguish and to puzzle together into longer stretches; but the bases in nucleic acids are only four. By the late 1960s, just when the crisis in the molecular biology of eukaryotes was becoming painful, Sanger was working with RNA, single-stranded, from viruses. As the technology of recombinant DNA became available, however, he switched to DNA.

He chose to work with the DNA of the bacteriophage known as ØX174 ("phi-X 174")--a shrewd choice, for the virus is one of the tiniest, with few genes and little DNA. Its DNA is not a double helix but a single strand, in the form of a circle. It has a total, we now know, of 5,375 nucleotides. In contrast, the two chains of the insulin molecule total 51 amino acids. None the less, ØX174 gave Sanger a source of identical molecules, short ones as DNA goes, in quantity.

The difficulties were formidable. Sanger had sequenced protein by breaking the chains in a few fixed places with specific digestive enzymes, segregating the fragments chromatographically into homogeneous sets, breaking each fragment into subsets, and so on. Repeated degradation and separation of fragments and subfragments of the ØX174 DNA, though, demanded immense labor and often gave as end product materials too scant to work with.

His solution was brilliant and elegant. Instead of breaking down DNA, he grew it. Starting with a supply of ØX DNA, he set up a system for replicating, a solution containing the necessary components including the four nucleotides with one of them labelled radioactively. To this he added a DNA polymerase and let the system string together complementary DNA for a limited time. The polymerase started replication at different moments, but always at the same initiation site. He filtered the resulting mixture to get rid of remaining free nucleotides. He was left with a mixture of DNA strands of all different lengths each begun at the same fixed point but some ending one base later, some two, some three, and so on, up to about 80--and labelled radioactively.

In principle, these short polymers of different lengths could be separated by a technique, a variant of chromatography, called electrophoresis, in which a weak electric current causes the molecules in solution to migrate across a slab of gelatine, longer chains moving faster and so farther. Gel electrophoresis is indeed sensitive enough to discriminate pretty accurately between small groups of molecules differing in length by as little as one nucleotide. That would yield a series of bands, one for each length, detectable by their radioactivity when a sheet of photographic paper was laid on the slab. But what Sanger wanted to know was the identity of the nucleotide--its base--at the tip of each strand. This the single series of bands would not reveal.

As the next step, therefore, Sanger reincubated samples of the mixture with DNA polymerase under various conditions. In what he called the "minus" system, the reincubation takes place in a solution containing only three of the four types of nucleotides. The polymerase makes all the strands start growing again, adding nucleotides one at a time according to the ØX base sequence. But when a strand reaches a place where the ØX sequence calls for the missing nucleotide, growth stops. If the missing nucleotide is A, the result is a mixture where every chain, whatever its length, requires A as its next base. Similar incubations minus T, minus G, and minus C produce, in all, a set of four samples, each a mixture of lengths where the identity of the next base, beyond the tip of the strand, is known. When the four samples are fractionated by electrophoresis in parallel on the same gel, the result is "a pattern of bands" in not one but four tracks, Sanger said in 1975, "from which, under ideal conditions, the nucleotide sequence can be read off."

[ILLUSTRATION HERE]

To confirm the results, "since ideal conditions are not often obtained," Sanger devised the "plus" system. Once again, samples of the original mixture of polymers of all different lengths are reincubated with DNA polymerase. The polymerase this time, though, is the product of a bacterial virus called T4, and has an unusual property. In a solution containing the necessary nucleotides, this enzyme actively knits DNA strands. But if nucleotides are missing, the enzyme slowly breaks down any DNA present, starting at the end where it would normally be lengthening the chain, and instead nibbling nucleotides away one at a time. Thus, in a solution where A nucleotides are the only ones present, the enzyme degrades each strand, nucleotide by nucleotide, until it reaches an A. At this point, enzyme molecules replace those As faster than others can take them off. The result is a mixture where every strand, whatever its length, has the base A at its tip. Similar incubation plus T, plus G, and plus C produce, once more, a set of four samples that, on electrophoresis, yield a pattern of bands that read out the sequence directly.[40]

A baffling mystery was reduced to a graphic pair of charts. Sanger and colleagues had published portions of the work over several years, but he presented the full system as part of a lecture at the Royal Society in London, on 15 May 1975. The auditorium there is large, but was packed, scientists standing along the walls, sitting in the aisles. The portrait of Charles II, the society's founding patron, looked down from behind the lectern. When Sanger put up a slide showing all eight bands of the plus-and-minus method, run on the same gel, and began reading off the base sequence, the audience sat immobile. He then showed that the base sequences of the portions of the viral genome he had been dealing with were confirmed by the amino-acid sequences which were known for two viral proteins. When he reached the end of the lecture the audience rose to applaud.

During that same period, Walter Gilbert and A. M. Maxam, at Harvard, devised a method for sequencing DNA using chemicals rather than enzymes, but with comparable simplicity and directness. Since then, the two methods have been standardized, speeded up, in large part automated: a single technician can now sequence hundreds of base pairs a day.[41]

Early in 1977, in a set of papers in Nature, Sanger and his group published the full base sequence, all 5,375 nucleotides, of ΦX174. In sequencing the genome, they had mapped it. The difference is instructive--and the map had startling features. The ΦX174 genome comprises nine structural genes--coding, that is, for nine different proteins--plus control elements. These controls included the site where replication of the viral DNA begins, and for each gene the points where transcription starts and stops. The most spectacular discovery was that certain stretches of this virus's DNA in fact overlap genes for two proteins simultaneously. In these stretches, the amino-acid sequence of one protein is encoded starting at one site on the DNA, while the other starts farther along and shifts the reading frame so that the triplets dictate an entirely different series of amino acids--the first gene reading codons as 1,2,3 4,5,6 7,8,9 10,... and the second reading them 3,4,5 6,7,8 and so on.[42]

The anatomy of the human genome

The first conference on the mapping of human genes took place in New Haven in 1973. Until shortly before that time, human genetics had hardly progressed beyond the sort of observation available to Garrod. Even in a large family, the number of children is small for genetical analysis; geneticists cannot, of course, conduct human breeding experiments. An increasing number of human characters were known to be inherited as Mendelian genes. Most of these were disorders caused by biochemical deficiencies. Even to associate these in linkage groups, other than the group on the X chromosome, was difficult: as late as 1967, only nine human linkages had been established, and seven of those were mere pairs of genes, one was a triplet, and one a quadruplet. To assign even such rudimentary linkage groups to specific human chromosomes was impossible. Indeed, not until 1956 did cytologists get the number of human chromosomes right. For decades, the count stood at 48--23 pairs of non-sex chromosomes (called autosomes) and the X-Y pair. Techniques for squashing mammalian cell nuclei to spread the chromosomes out for photomicrography improved only gradually; in 1956, [fisrt name] Tjio announced that the correct human chromosome number is 46--comprising 22 pairs of autosomes, X, and Y. But to identify particular human chromosomes with assurance could not be done. For example, chromosomes 6 through

12, as now numbered, and the X chromosome appeared virtually identical under the microscope.

In 1967, Mary Weiss and Howard Green took a crucial first step in human-gene mapping with the publication of a technique for fusing human cells and mouse cells growing in culture. The technique depends on a mammalian-tumor virus called Sendai virus. A virus necessarily has a site on its surface with which it attaches to a host cell in order to inject its nucleic acid. The peculiarity of Sendai virus is that each particle has several attachment sites. Added to a culture, a Sendai particle will attach to several cells; far smaller than the cells, it necessarily holds them in close contact. Weiss and Green showed that appropriate biochemicals will soften the cell walls; some cells then fuse; some of the hybrids survive and multiply. Though the hybrids tend to throw off the human chromosomes as they go through successive divisions, lines of hybrids can be stabilized reproducing the full mouse genome--on twenty chromosomes--and a few human chromosomes as well. Mouse and human chromosomes are easily distinguished under the microscope; the protein products of an increasing number of human genes are known, of course, and can be detected in the biochemical production of hybrid cells. Mapping deductions were possible. Thus even in that first paper Weiss and Green observed that the enzyme thymidine kinase was related to a specific human chromosome--but which of the several in the hybrid cell, and overall in the human chromosome set, they could not say. The technique was obviously of immense promise for genetic analysis. The geneticist Guido Pontecorvo called it "parasexual." J. B. S. Haldane called it "an alternative to sex."[43]

In 1970, Torbjörn Caspersson, L. Zech, and colleagues, in Sweden, published the first method for staining human chromosomes in such a way that banding patterns appear, like those Painter had found in the giant chromosomes of fruit flies nearly forty years earlier. Caspersson has been working with locating and identifying nucleic acids since the early 1930s, when his doctoral dissertation presented a combination of spectroscopy and microscopy that could find and follow nucleic acids in the living cell. To band human or other mammalian chromosomes, Caspersson used fluorescent dyes, chief among them quinacrine. These required an ultraviolet microscope and were difficult to work with. But within a year, several other laboratories came up with staining methods that did not require special instrumentation. The banding technique most widespread today uses a dye called giemsa.[44]

These methods revealed a physical anatomy of the human (or other mammalian) chromosome set. They brought out patterns of bands, thick and thin, that make each human chromosome unambiguously identifiable. One could now pair them up by matching stripes and distinguish them from others in the cell--the two of chromosome 6, say, obviously different from the two of chromosome 7, and so on. One could pick out any particular chromosome--11, say--in cells from different individuals. One could spot changes in individual chromosomes--inversions, translocations, deletions, duplications--that alter the banding pattern. One could tell exactly which human chromosomes remain in a stable line of mouse-human hybrid cells.

Somatic-cell hybridization took off when banding became available, growing logarithmically from 1971. The first human-gene mapping conference, in 1973, was based for the most part on the rapid development of mapping by somatic-cell hybridization which had taken place in just a couple of years. Geneticists began to assign genes to specific chromosomes and even to subregions of the chromosome--sometimes to specific bands. To do this, they used as the human component in a cell-fusion experiment a cell line in which the usual chromosomes were altered in some way, say by reciprocal translocations. The techniques, though complex in detail, were in concept the direct elaboration of the classical gene mapping of Morgan, Sturtevant, McClintock, and Painter. Banding and somatic-cell hybridization made linkage and crossover studies fruitful with human genes.[45]

The anatomy of the genetic material is exploited at a finer scale, at the level of the molecular sequences of DNA itself, by a method developed in the late 1970s that combines close analysis of pedigrees with the most basic tool of the genetic engineers. So long as the study of the transmission of characters within families depended on tracing variations in the gene products--the enzymes, ultimately the discernible characters, and their mutations--human genetics could not progress much beyond Garrod. To be sure, sophisticated statistical techniques were developed to squeeze more out of the meagre data, yet the practical results were limited. Then restriction enzymes of high specifity became available, which cut DNA precisely at positions defined by sequences of bases--for example, at the downstream end of every sequence ATTGTCA and nowhere else. Such an enzyme, as we've noted, breaks a given sample of DNA into fragments of varying length.

By 1978 and 1979, studies had begun to find that when a restriction enzyme is applied to DNA from different individuals, the resulting sets of fragments sometimes differ markedly from one person to the next. The reason had to be, of course, that the presence of the restriction sites--the specific sequences the particular enzyme recognizes--varies from one to another, due to mutations in the base sequences. This offered the chance to get at variation between individuals not at the level of the character or the enzyme but at the level of the genetic material. The odds would be against the variations hitting within a particular gene of interest; but several geneticists saw at once that variations anywhere in the neighborhood of a gene could usefully serve as markers. Applied to portions of DNA from a number of different members of a family afflicted with a genetic disorder, the method might identify individuals who carry a recessive gene for the disorder, or who actually have the disorder but have not yet manifested it.

The variations in DNA are called restriction-fragment length polymorphisms, or RFLPs (pronounced "riflips"), and their study is already proving immensely productive. Several instances of great clinical promise have gained press attention recently, beginning in 1983 with publication of a large investigation of an extended family in Venezuela that is scourged by Huntington's chorea. This disease is a progressive nervous degeneration, always fatal. It is caused by a single Mendelian gene that is dominant, which means that any person with the gene will inevitably come down with the disease, and that any offspring will each have one chance in two of getting it. However, the product--enzyme or whatever--of the mutant gene or of its normal counterpart is unknown. And the disorder shows up relatively late in life, usually in the mid-forties or beyond. The combination of factors is alarming, because individuals have often had children before coming down with the disease. The Venezuelan study, by James Gusella and a small army of colleagues, demonstrated that members of the family with the disease showed a characteristic and distinct pattern of restriction-fragment lengths. Some other family members in whom the disease had not begun its course also showed the characteristic pattern--and these eventually came down with it.[46] Since that study, the same methods have yielded results with cystic fibrosis, adult polycystic kidney disease, Duchenne muscular dystrophy, and others. In many instances the method has located the gene on a specific chromosome. The next step, geneticists hope, will be to isolate the gene itself for such a disorder, sequence it, and thus identify the gene's product--and so arrive at its mode of action and thus at the possibility of a therapy. The approach has been called reverse genetics.

The limitation in the use of RFLP markers for clinical work is that they must be found and analyzed anew for each study. A family with Huntington's disease in North America will show DNA polymorphisms, but not the same ones as found in the Venezuelan cases.

Beyond the clinical applications, some geneticists, prominent among them David Botstein at the Massachusetts Institute of Technology, have argued since RFLPS first came into use that they will play a major part in mapping and sequencing the human genome.

Another big technical step was taken in 1980, with what is called in situ hybridization of genetic material--and this technique is altogether new, with no obvious precedent in classical genetics. The approach is direct, though it requires high technical precision. The trick begins with a length of DNA that contains a gene or even a short strech of a gene; often, for example, the geneticist will start with a piece of messenger RNA and use a system containing a reverse transcriptase to read that back into DNA. This is then cloned in bacteria, to grow up a suitable quantity. These methods are taken over directly from the genetic engineers. (The first human genes to be cloned were two, one for the beta chains of hemoglobin and the other for a hormone called chorionic somatomammotropin; the work was done in two different laboratories essentially simultaneously and published in 1977 [check], and the priority between them is uncertain.)

For in situ hybridization, the cloned DNA is grown with a radioactive label. The supply of labelled genetic material is called a probe. The geneticist then prepares a spread of the chromosomes of a human cell, treated so that the DNA of those chromosomes can form base pairs with matching sequences of DNA. He applies the probe to the chromosome spread. The probe will form hybrids with the cellular DNA but only at the place where the base sequences match up. The geneticist then makes an an autoradiograph of this, producing a sheet of film, a picture of the chromosomes, with black dots due to the radioactivity at the points where the genes are located. The idea had been tried by many. It worked well with genes that are present on the chromosome in multiple copies. But the genes of greatest interest are likely to offer just one copy in the genome, and with them the amount of radioactivity was too low to show up. Several experimenters got erroneous results. But in 1980, Mary Harper, WHERE, added one ingredient to the technique. She found that dextran, which is a carbohydrate built up of glucose molecules, when added with the probe forms a mesh or tangle that makes the molecules of the probe cluster; as a result, the hybridization site has enough radioactivity to show up clearly. In this way, in 1981, Harper and two colleagues mapped the gene for insulin. She prepared an autoradiograph that showed a black blob from the radioactivity at the tip of the short arm of chromosome 11.[47] That year, mapping by in situ hybridization became a standard method.

Thus today, whenever a gene gets cloned the genetiticist has not completed its characterization until he determines which chromosome is carrying it and where. The gene is first run through a panel of somatic-cell hybrids. When it is found to be on a particular chromosome, say 12, then the geneticist confirms the assignment and narrows it down to a specific part of that chromosome by in-situ hybridization. One great advantage is that the gene does not have to be actively expressed in the cell. The geneticist finds the gene directly, not by its products.

Conclusion: the sequences and the maps

The technology of genetic experiment and analysis has done more than facilitate research and theory. It has driven research and theory--nowhere more obviously than in the changing concepts of the gene, the sequence of genes, and the map of genes. For Morgan and Sturtevant, the map and the sequence were virtually the same, a linear array of elements spaced in proportion to crossover frequencies. Today, the conceptual difference between genetic sequences and maps is great and all important.

The sequence is now simply defined and directly accessible: it is the order of base pairs in the DNA from the long arm of chromosome 1 to the short arm of X. The sequence has one important conceptual complication, which is that humans are enormously diverse, genetically, more diverse by far than any other species. No one sequence can be established as the type or norm. Parallel sequences--for the most interesting genes and families of genes, many parallel sequences--will be needed. None the less, the technology is already available and is rapidly getting better: this technology, like, say, transistors in the early days, is still in that pleasant phase when every major new application of it entrains great reductions in unit costs.

The sequence is now, in fact, the territory. But maps are never the territory mapped. They are at once less than the territory and more. They are abstractions from the territory, leaving much out; and they are labelled, the features of the terrain identified. Every natural territory generates many different maps, and the differences spring from the purposes to which the maps are to be put. The political map differs from the road map which differs from the map of natural resources as seen from mapping satellites, and so on. Yet all maps share one general characteristic: the abstracting and the labelling establish a set of relationships. In short, by omitting much of the territory maps gain explanatory power. By being less than the sequences, the maps of the genome will be far more.

About a thousand genes have been mapped to chromosomes, which amounts to nearly a fourth of all the genes that McKusick lists with any confidence in the catalogue of Mendelian inheritance in man. Present estimates place the total of human genes between 50,000 and 100,000--and this range has remained remarkably stable under different methods of estimating over several decades. The best present estimate depends on the density of structural genes along the most intensively studied piece of chromosome, the short arm of 11, which contains besides the insulin gene a stretch coding for several different ersions of the beta chain of hemoglobin--versions that are expressed in slightly different proteins in embryonic, fetal, and adult life. The globin genes are known to be relatively small and densely spaced; when those factors are allowed for, the total bulk of the chromosomes yields the range, once more, of 50,000 to 100,000 genes. Even at the lower figure, an enormous portion of the human genome is as yet unknown to us: for 40,000 genes or more we don't even know of mutations that reveal them by their variants or defects.

The sequence by itself is dry and uninformative--an all-but-endless string of A, T, G, C, in no predictable order. The map is going to be more complex than we can yet imagine. The contrast is rooted in a single cause: the human genome was not designed, it evolved. The relationship of the map of genes to the living creature will have to be analyzed almost instance by instance.

Even at the simplest level, the locations of structural genes, mapping has progressed far enough to establish certain daunting negatives. Genes for one organ or one system or one function do not confine themselves to some one chromosome. Genes for the enzymes for the several steps in a biochemical pathway are not to be found lying together in an operon as in the bacterial model, but will usually be on different chromosomes. Even where a protein, like hemoglobin, is made up of several different chains, almost without exception the genes for the different chains are not on the same chromosome. Thus, the gene for the alpha chain of hemoglobin is coded by chromosome 16.

At the next level, the map must account for regulation, for genetic sequences that are not structural but control elements. This will include transposons--jumping genes--as well as, necessarily, whole classes of controls we have hardly glimpsed. The map must also sort out the problem of the immense amount of DNA that has no function we know, or that may be functionless. In doing so, it will have to relate the various segments of structural genes broken up by intervening sequences, and show how the genes get put back together again.

The penultimate level of the maps is the developmental: what we want the map to show us is what gets turned on and off in temporal sequence during the human creature's lifetime, to answer the intractable questions of embryology, of differentiation. Special versions of the developmental map will be clinical, dealing, say, with the inborn errors of metabolism and their cure by gene therapy, or with the cancers and their prevention.

The ultimate map will extract from the sequences and from all the preceding levels of mapping, like the layers of an archeological dig, whatever suggestions about the evolution of man the steps of that evolution have not obliterated.

Notes

1. United States Congress, Office of Technology Assessment, *Human Gene Therapy: Background Paper* (Washington, D. C.: 1984)
2. Darwin, Charles, *The Variation of Animals and Plants under Domestication* (London: John Murray, 1868), *passim.*, but particularly vo. 2, chapters 12, 13, and 27
3. Mendel, Gregor, *Versuche über Pflanzen-Hybriden, Verhandlungen des Naturforschenden Vereines in Brunn* 4 (1865): 3-47. Best English translation is by Eva R. Sherwood, in Curt Stern and Eva R. Sherwood, eds, *The Origin of Genetics* (San Francisco: W. H. Freeman, 1966), pages 1-48.
4. See the letter from Francis Galton to Charles Darwin, 19 December 1875, cited and quoted in Robert C. Olby, *Origins of Mendelism* (New York: Schocken books, 1966), pages 71-72.
5. Weismann, August, *The Germ-Plasm: A Theory of Heredity*, tr. by W. Newton Parker and Harriet Rönnfeldt (London: Walter Scott, Ltd, 1893), especially pages 1-35
6. Garrod, Archibald E., "The Incidence of Alkaptonuria: A Study in Chemical Individuality," *The Lancet* 2 (13 December 1902): 1617-1620; *idem*, *Inborn Errors of Metabolism* (London: Frowde & Hodder, 1909)
7. McKusick, Victor, *Mendelian Inheritance in Man*, seventh edition (Baltimore: Johns Hopkins University Press, 1986)
8. Sutton, Walter S., "The Chromosomes in Heredity," *Biological Bulletin* 4 (1903): 231-251
9. Bateson, William, and R. C. Punnett, "Experimental Studies in the Physiology of Heredity," *Reports to the Evolution Committee of the Roayal Society*, Reports 2, 3, and 4 (1905-1908), reprinted in part in James A. Peters, ed., *Classic Papers in Genetics* (Englewood Cliffs, N. J.: Prentice-Hall, 1959), pages 42-60
10. East, Edward M., "A Mendelian Interpretation of Variation That Is Apparently Continuous," *American Naturalist* 44 (February 1910): pages 65-82
11. Morgan, T. H., "Sex Limited Inheritance in Drosophila," *Science* 32 (1910): 120-122
12. Wilson, E. B., "The Sex Chromosomes," *Arch. Mikrosk. Anat. Entwicklungsmech.* 77 (1911): 249; cited in McKusick, Victor, "The Gene Map of *Homo sapiens*: Status and Prospectus," *Cold Spring Harbor Symposia on Quantitative Biology* 51 (1986): 15, 24.

13. McKusick, *Mendelian Inheritance in Man, passim.*

14. Sturtevant, A. H., "The Linear Arrangement of Six Sex-linked Factors in Drosophila, as Shown by Their Mode of Association," *Journal of Experimental Zoology* 14 (1913): 43-59

15. See, for example, Morgan, T. H., A. H. Sturtevant, H. J. Muller, and C. B. Bridges, *The Mechanism of Mendelian Heredity*, revised ed. (New York: Henry Holt, 1922), frontispiece, p. ii.

16. Sturtevant, A. H., "The Effects of Unequal Crossing Over at the Bar Locus in Drosophila," *Genetics* 10 (1925): 117-147

17. McKusick, "Gene Map of *Homo sapiens*," 19

18. Muller, H. J., "Variation Due to Change in the Individual Gene," *American Naturalist* 56 (1922): 32-50

19. Muller, H. J., "Artificial Transmutation of the Gene," *Science* 66 (1927): 84-87

20. Creighton, Harriet B., and Barbara McClintock, "A Correlation of Cytological and Genetical Crossing-Over in *Zea mays*," *Proceedings of the National Academy of Sciences* 17 (1931): 492-497

21. Painter, T. S., "A New Method for the Study of Chromosome Rearrangements and Plotting of Chromosome Maps," *Science* 78 (1933): 585-586

22. Conversation, Max Delbrück and Horace Freeland Judson, 9 July 1972, quoted and cited in Judson, Horace Freeland, *The Eighth Day of Creation* (New York: Simon & Schuster, 1979), pages 58-60.

23. Crick, F. H. C., "On Protein Synthesis," *Symposia of the Society for Experimental Biology* 12 (1957): 137

24. Wright, Sewall, "Color Inheritance in Mammals," *Journal of Heredity* 8 (1917): 224-235; *idem*., "The Physiology of the Gene," *Physiological Reviews* 21 (1941): 487-527

25. Provine, William B., *Sewall Wright and Evolutionary Biology* (Chicago: University of Chicago Press, 1986).

26. Judson, *The Eighth Day of Creation*, pages 214, 356, 368

27. Discussed at greater length, with citations, in Judson, Horace Freeland, "Reflections on the Historiography of Molecular Biology," *Minerva* 18 (Autumn 1980): 369-421.

28. Sanger, Frederick, "The Terminal Peptides of Insulin," *Biochemical Journal* 45 (1949): 563-574, quote at 573; *idem*, "Some Chemical Investigations on the Structure of

Insulin," *Cold Spring Harbor Symposia on Quantitative Biology, 14: Amino Acids and Proteins* (meeting 8-16 June 1949): 153-160

29. Chargaff, Erwin, "Chemical Specificity of Nucleic Acids and Mechanism of Their Enzymatic Degradation," *Experientia* 6 (1950): 201-209; for other papers and discussion, see Judson, "Historiography."

30. Luria, S. E., and M. Delbrück, "Mutations of Bacteria from Virus Sensitivity to Virus Resistance," *Genetics* 28 (1943): 498-511. Monod, Jacques, et Alice Audureau, "Mutation et adaptation enzymatique che Escherichia coli-mutabile," *Annales de l'Institut Pasteur, Paris* 72 (1946): 868-878; published late because of wartime restrictions.

31. Judson, *The Eighth Day of Creation*, pages 19-195

32. Jacob, François, and Elie L. Wollman, "Induction of Phage Development in Lysogenic Bacteria," *Cold Spring Harbor Symposia on Quantitative Biology* 18 (1953): 101-121; Hayes, William, "The Mechanism of Genetic Recombination in *Escherichia coli*," ibid., pages 75-93; Wollman, Elie L., and François Jacob, "Sur le mécanisme du transfert de matériel génétique au cours de la recombinaison chez *Escherichia coli* K12," *Comptes rendue Academie Sciences 240* (séance du 20 Juin 1955): 2449-2451; interviews with François Jacob, 3 December 1975 and 28 May 1977

33. The best contemporary summary of the early work is Jacob, François, and Elie Wollman, "Genetic and Physical Determination of Chromosomal Segments in *Escherichia coli*," *Symposium of the Society for Experimental Biology* 12 (1958; held in 1957): 75-92.

34. Benzer, Seymour, "The Elementary Units of Heredity," in McElroy, W. D., and Bentley Glass, eds., *The Chemical Basis of Heredity* (Baltimore: Johns Hopkins University Press, 1957): 70-93

35. McClintock, Barbara, The Origin and Behavior of Mutable Loci in Maize," *Proceedings of the National Academy of Sciences* 36 (1950): 344-355.

36. The discovery and naming of the operon and the repressor were done in an intricate series of small experimental steps; for thorough discussion and citations, Judson, *The Eighth Day of Creation*, pages 400-435

37. Jacob, François, and Jacques Monod, "Genetic Regulatory Mechanisms in the Synthesis of Proteins," *Journal of Molecular Biology* 3 (1961): 318-356

38. Nirenberg, M. W., and J. H. Matthaei, "The Dependence of Cell-Free Protein Synthesis in *E. coli* upon naturally occurring or synthetic polynucleotides," *Proceedings of the National Academy of Sciences* 47 (1961): 1588-1602

39. Crick, F. H. C., conversation, 15 September 1970; cited in Judson, *Eighth Day of Creation*, pages 204, 640.

40. For the history of the discovery, with full citations, see Temin, Howard M., Nobel Lecture, 12 December 1975, published as "The DNA Provirus Hypothesis," *Science* (192): 1075-1079.

41. Sanger, Frederick, "The Croonian Lecture, 1975: Nucleotide Sequences in DNA," *Proceedings of the Royal Society* of London *B* 191 (1975): 317-333; quotes at 326.

42. Maxam, A. M., and W. Gilbert, "A New Method for Sequencing DNA," *Proceedings of the National Academy of Sciences* 74 (1977): 1258

43. Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III, P. M. Slocombe, and M. Smith, "Nucleotide Sequence of Bacteriophage ØX174 DNA," *Nature* 265 (24 February 1977): 687-695; see also the three following papers in that issue, pages 695-705.

44. Review of all this work with extensive citations: McKusick, "Gene Map of *Homo sapiens*"

45. Interview, Victor McKusick, 6 June 1987

46. McKusick, "Gene Map of *Homo sapiens*"; interview, McKusick, 6 June 1987

47. A review is: Gusella, J. F., *et al*.,"Molecular Genetics of Huntington's Disease," *Cold Spring Harbor Symposia on Quantitative Biology* 51 (1986): 359-364.

48. Interview, McKusick, 6 June 1987

D[illegible] 1 1987

Long-Term Implications of Mapping & Sequencing the Human Genome

Ethical and Philosophical Implications

Marc Lappé, PhD

University of Illinois at Chicago
Humanistic Studies Program
Center for Educational Development
College of Medicine

TABLE OF CONTENTS

CHAPTER PAGE

1 INTRODUCTION 1
  a. First order implications 2
  b. Second order implications 6
  c. The context of research 7

2 BROAD PHILOSOPHICAL CONSIDERATIONS 9
  a. Value-neutrality 9
  b. Deontological concerns 10
  c. Consequentialist arguments 14
  d. Conclusions 16

3 MAJOR PRINCIPLES OF ETHICS 17
  a. Autonomy 17
  b. Beneficence 20
  c. Justice 23
  d. Intergenerational responsibility 29

4 MAJOR ARGUMENTS FOR EXPANDING EXISTING MAPPING/SEQUENCING PROGRAMS 31
  a. Issues of scale 31
  b. Arguments for high resolution mapping 33
  c. Technical readiness 33
  d. Vaues for medical science 34
  e. Intrinsic values 34
  f. Counterarguments 35
  g. Spinoffs 35

5 LONG-TERM IMPLICATIONS 37
  a. Level 1 considerations 38
  b. Level 2 considerations 40
  c. Level 3 considerations 40

6 EUGENIC IMPLICATIONS 44
  a. Judeo-christian traditions 46
  b. Positive eugenics 47
  c. Negative eugenics 50
  d. Nefarious uses 51
  e. "Normal" eugenics 52
  f. Voluntarism 54
  g. Privacy 54

7 ISSUES OF CONTROL OF KNOWLEDGE 55
  a. Options: The options of the government appear at least threefold 57
  b. Public interest 58
  c. Arguments 59

8 SUMMARY AND RECOMMENDATIONS 60

REFERENCES 62

Long-Term Implications of Mapping & Sequencing the Human Genome
Ethical and Philosophical Implications

Marc Lappé, PhD
University of Illinois at Chicago
College of Medicine

1. Introduction

A project to map and sequence the entire human genome would be the most ambitious undertaking in human biology. On completion, the exact location and sequence of the nucleotides that make up the 3 billion units of human DNA would be assembled, catalogued, and maintained (1). Ethical issues will arise in the course of deciding what priority to give the project and what level of resolution (broad map, linkage group picture, full sequence) is most appropriate given our current level of ability. More profound ethical questions will be posed by the possible applications of the data as genetic data reveal more about the underlying basis for human disease, ability and social interaction. Questions about personal liberty, privacy, "right to know" and societal versus individual prerogatives to use or sequester genetic information are among the most important.

A full genomic picture will of necessity raise anew the desirability of using genetic knowledge and its control to shape the future of human society. The complexity and urgency of resolving these and related issues will increase in proportion to the stage of development of the project.

a. First Order Implications

The hierarchy of ethical issues thus begins in the realm of the possible and extends into more futuristic applications of genetic knowledge. The most immediate question is that of cost-effectiveness and priority-setting. The genome project is presently costly, and will require substantial investment in the time and efforts of researchers currently engaged in more directly innovative or creative research. At a minimum, the reagent cost per base uncovered is currently $0.015 (2), so that a 3 billion sequence would cost $45 M just for a single "run" in chemicals alone. Replication of runs to assure accuracy, maintenance costs (estimated at $250M), and personnel and technical equipment costs, will undoubtedly bring the total to at least one half billion dollars.

The ethical issue of the possible deferral of projects that might more directly assist public health to pay this cost must be balanced against the possible long-term gains from this new repository of data. In comparison to the cost outlays for military projects, such as the B-1 Bomber, the outlay of a half or even several billion dollars for a permanently useful data base suggests that priority be given to this project. Deferral of the project until less expensive technologies arrive appears unwarranted, since at least two reasonably priced (@ $90,000 each) sequencers are now available. Proficiency in their use will undoubtedly improve with effort, while useful data are being assembled.

The arguments for acquiring the sequence data now could include the lost opportunity to diagnose and/or develop treatments for diseases which have already been shown to have a genetic base. In aggregate, these disorders comprise the single most important source of pediatric hospitalizations (if juvenile onset diabetes is included) (3), and individually account for substantial human suffering and disability. Similar "costs" of not initiating the project would accrue to the research community which has a pressing need for data on which to derive functional relationships between genes and their polypeptide sequences, protein structure and configuration, and precise analyses of the evolutionary relationships and distances between living organisms.

The countervailing argument could emphasize that in the first phase of work, the genome project will generate enormous volumes of data of uncertain significance. According to this line of reasoning, if we understood the intrinsic structure of the genetic array first (e.g. by developing a full-scale map of the genome prior to a full sequencing project), we might avoid unnecessary acquisition of "junk" DNA or insignificant repititious sequences.

Such a preliminary exercise might also provide time to channel the necessary education of public expectations to the rather limited successes that a full-blown project will initially engender. Currently, the public is likely to anticipate that the genome project will generate great discoveries with immediate medical and societal significance. Such an expectation is unlikely to be met simply because most of the sequence data initially will be unrelated to linkage groups of known significance. In the past, difficulties were encountered where genetic information was wrongly or mistakenly interpreted, such as when markers thought to be closely linked to disease loci were found to have too high a recombination frequency to be reliable. This problem occured recently in the research on the genetic link to Alzheimer's disease. Here, the linkage of a gene for amyloid protein and the characteristic pathology of both Alzheimer's and Down syndrome was mistakenly taken to imply that gene doubling could account for a causal association (4).

Similar problems can arise when the presence of the gene is a risk factor, but needs significant environmental interaction to function pathologically. In both instances, the uncovery of the presence of a gene will not mean an absolute certainty of disease, but the public's association of a "genetic" diagnosis with certitude may color the interpretation of the first data. An educational program to anticipate this likelihood would be extremely valuable.

The problem of unrealistic expectations is thus one that Congress and appropriate funding agencies should be aware of. The scientific community has only recently recognized the delay between acquisition of data and its association with recognizable structure and function relationships. This was particularly evident in the instance of the much heralded sequencing effort to define the Epstein Barr virus genotype. In this instance, the full genetic code has not yet provided the anticipated insights into the relationship between gene sequences and specific viral functions (5).

The value of genomic data will also initially be limited by the intrinsic problems of extrapolating genetic data to human biology generally or to population wide disease patterns specifically. Difficulties in interpreting genomic data will almost certainly occur in extending the information from gene sequences which are linked with human developmental controls (the so-called "homeo box"); associated with human proclivities to disease; or appear linked to human attributes like intelligence. Other discoveries will build on data already available that relate certain genetic loci and chromosomal regions to susceptibility to human disease, e.g., the HLA regions that control immune functions and histocompatility (6).

The most immediate applications will likely be those that fit into pre-existing technologies like prenatal diagnosis that depend on accurate genetic diagnosis. Even in its earliest stages, the genome project will expand the number and type of infirmities that researchers can identify prenatally. These data have already influenced how we identify and treat genetically handicapped persons or fetuses, and soon will influence how we attempt to remedy the expression of flawed human genes (cf 7). The opportunities for adopting powerful and socially disruptive policies based on this data are greatest in this area of early development. It is highly likely, for instance, that new genetic technologies will find early application within the public policy window created by permissive or absent policies towards artificial insemination, embryo transfer, and selective abortion.

While most of these problems turn on issues already being discussed in public policy forums, the same familiarity will not apply to applications of genetic data to adults. For instance, genes uncovered by sequencing will almost certainly include some that cause or contribute to increased proclivities towards serious debilitating but chronic adult disorders, including Huntington's disease, Alzheimer's disease and multiple sclerosis (8). These data will raise new questions of competence, dependency and human autonomy, especially in the choice of whether or not to seek out the genetic knowledge and how to use it. Other genes or markers will likely contribute to the growing body of information about susceptibility to occupational stresses, hazardous substances or infectious agents, including the human immunodeficiency virus (cf 9).

These data impact on the growing public movement towards universal "right to know" laws and potentially conflict with traditional norms of confidentiality. The ultimate questions in the human genome project are thus analogous to those encountered with technological inventions in the rest of society: the resolution of individual versus community rights; societal versus personal prerogatives; and, the distribution of valuable technical resources.

b. Second Order Implications

The more global social and political implications of gene sequencing are likely to emerge as new technologies spawned by control of sequence data begin to impact on the industrial and commercial sectors. It is here that the political issues of the nature and degree of control by persons, governments or the private sector over procreative decisions by individuals with specific genotypes, and possible limitations on their employment, insurability, and liberties loom largest.

In these penultimate applications, the human genome project raises profound philosophical questions about the nature of human freedom, the operation of autonomy, and the definition of humanhood. To some, like Jeremy Rifkin, such potentialities suggest an Orwellian world in which power belongs to a few technocrats who would reshape the evolutionary picture of our species (10). While these uses are currently remote and limited by legal and societal strictures on the compulsory use of genetic knowledge in screening or counseling settings (11), these issues are highly charged in the public mind (cf 12). Hence, it is important to provide guidelines for the implementation of gene sequence data that recognize ethical concerns early on.

The magnitude of the investment in the project also dictates that the issue of distribution of the costs and benefits of the project be resolved before more of the science of the human genome leaves the laboratory to the clinic and corporate world. These questions in turn heighten the urgency for resolving the issue of public versus private investment in the search for sequences, since it is at this level that the later questions of access to the resulting knowledge and ownership of the generated data will largely be determined.

c. The Context of Research

A fundamental concern about new knowledge is how it may reinforce existing societal structures or institutions that stereotype or stigmatize certain groups or otherwise reduce their freedoms and rights. Different degrees of caution, concern and regulation are appropriate for different political climates. For instance, strong arguments against embarking on a genome project could be made if the knowledge were likely to be abused in a totalitarian regime (eg, in 1928-1945 Germany [cf 13,14]); if the prospect for misuse (e.g., in eugenic programs (15) were high; or if a strong likelihood existed for using the newly acquired data to impose or reinforce racially discriminatory programs or policies as occurred in the early 1970s with genetic screening programs (16).

Historically, genetic data have been gathered to reinforce certain political ideologies (e.g., when Nazi studies attempted to confirm racist stereotypes) or to shape a national policy (e.g., when Lysenkoist views of the primacy of the environment over genetic inheritance were accepted as the sole policy base for Soviet agriculture in the 1950s). These distorted uses of genetic science had disastrous consequences (17) and warrant scrutiny for possible precedents in the genotype project.

One way to do this is to examine our own "Zeitgeist" to look for clues about pre-existing biases, stereotypes, or political theories to which genetic knowledge may be bent. Leon Kamin of Northeastern University, for instance, believes that we are in an era in which the public is much more ready to accept genetic explanations of complex human behaviors than was true just 10-20 years ago (18). We continue to have racial and homophobic tensions (e.g., those centering on AIDS) that suggest that the use of new genetic data about racial partitioning of biological attributes or the vulnerability of certain genotypes to costly diseases could be biased or discriminatory.

We are also beginning to see new studies in genetic journals (cf 19) that indicate a growing awareness and concern about the dysgenic consequences of present-day "laissez faire" genetic policies. Although geneticists have always known that some genetic policies (e.g., the selective abortion of fetuses with recessive disorders) could in the long run lead to an increase in the frequency of certain deleterious genes, it was always taken as a given that the function of human genetics was to alleviate human suffering and not to move towards more eugenic and hence coercive objectives. The presence of such an article does not necessarily signal a shift in that policy, but shows a readiness to recognize and perhaps act on longer term implications of genetic policies.

Two further elements of the present political climate warrant comment: First, we are in a period of great societal stress from the escalating costs of health care. These trends may encourage the early application of genetic data to maximize hospital efficiency and cost-effectiveness generally (e.g., through new diagnostic tests), or more ominously, the pre-emptive use of genetic testing to disqualify certain persons from insurance coverage or to limit their employability. Second, we are in an era in which substantial individual freedom exists over reproductive decision-making, and for which genetic data are avidly sought by some (notably in the higher socioeconomic brackets [cf 20]) to assist in deciding about the kind ("quality") of offspring to bear. While the overtly discriminatory policies for screening and identifying genetically handicapped or carrier (sickle cell trait) adult individuals appear to be issues of the past (21), the prospect for abuse of the new genome information must be anticipated within the area of procreative choices. A model for such possible misuse would come from studies which demonstrated the limited degree to which different segments of our society (but larger blocks of non-Western societies) have already used genetic data with amniocentesis or chorionic villus sampling to select against female fetuses--or newborns (cf 22).

These data indicate a climate of acceptance for using genetic data for reproductive decisions. Further erosion of the previous societal norm which permitted abortion only for extremely handicapping conditions has apparently recently occurred with the availability of tests for marginally impairing conditions such as protease deficiencies which predispose to early liver or chronic lung disease, and "mild" variants of metabolic disorders (e.g., phenylketonuria subtypes) (23). This group of circumstances suggests that eugenic applications of genetic data must be taken seriously.

2. Broad Philosophical Considerations

a. Value-Neutrality

The focus and scope of philosophical inquiry of the genome project depends to a substantial degree on whether the knowledge gathered is deemed neutral or value-laden. If genetic data are primarily value-neutral and akin to fundamental particle research in physics (at least in its data gathering phase), then the pertinent philosophical issues would be epistomological, such as whether or not the basic data can directly predict higher level phenomena. Some philosophers of science consider these latter issues of paramount importance (cf 24, 25), while others deem them trivial (26).

If the data are likely to affect human lives, rights and/or liberties in their foreseeable applications, then the project is assuredly not neutral, and the analysis should focus on ethical issues. Even a cursory review of the literature on contemporary human genetics shows that the overwhelming perception of applied DNA technologies is that they have powerful implications for human life (27-28). For this reason, this review will focus on the value-based, ethical issues raised by decoding the human genome.

b. Deontological Concerns

The fundamental question of whether or not the genome should be probed for its content precedes the question of what limits should be placed on what is found. An injunction to withhold acquisition of genome data would not come from a secular application of deontological principles. Nothing in a secular deontological ethic precludes the acquisition or incorporation of as much scientific knowledge into human experience (including ethical reflection) as possible, unless that incorporation entails violating a duty that holds for all rational persons.

The only place where such a duty might be derived is from certain forms of fundamentalist religious beliefs in which the mystery of human existence is considered sacrosanct. A duty not to investigate certain areas of the human experience could be inferred from a reading of Islamic texts, but the scope of this inquiry does not embrace perspectives outside of the American vein. As one example, Nasir Al Din Tusi, an Islamic philosopher, has observed that the reason for Islamic science's limited success in achieving comparable dominance over nature was rooted in concern over making the sacred profane (29).

The derivation of knowledge per se is not contraindicated by teachings in the Judeo-Christian tradition. In fact, most western religious traditions are silent on the issue of probing or uncovering the core genetic material. What is contravened, according to some canonical or Halachic teachings is the intentional manipulation of this genetic material (cf 30). This focus would impact on the uses of genomic information unveiled in the course of performing sequencing studies, but not on its acquisition. Members of both faiths resist the potential secularization of procreation predicated by the control of genetic material either at the level of selection of gametes (eg.g or sperm donation, AID, etc) or at the level of fetal genetic testing (cf 31).

The derivation of more knowledge which objectified the procreative act, could be seen as contraindicated by a naturalistic law which asserted that the continuity of the species is dependent on a perpetuation of "natural" (i.e., random) matings. This point can be rebutted by the simple observation that very few if any mating systems in human (or other primate societies) is strictly random. In fact mating is distinctly assortative (32), and increasingly linked to tacit or implicit recognition of ancestry, and by inferrence, genetics.

One major concern is that the genome project could increase our focus on the mechanistic aspects of living organisms rather than on higher order phenomena or relationships. The centerpoint of this concern in the Catholic tradition is on the sanctity of the conjugal act. Richard McCormick, representing one Catholic view, rejects donor insemination or embryo transfer on the grounds that it focuses on the mechanistic elements rather than the context of the conjugal relationship (33). Concern for this attenuation of respect for fundamentally human elements of the marital bond can be discerned in the Papal Letter on in vitro fertilization and embryo transfer (34).

In the Jewish faith, the elements of humanness are seen to be embedded in the genome such that special care and concern about and during procreation to ensure its propagation have been proposed (see Section 5, Eugenics). Nonetheless, spokespersons for the Talmudic tradition such as Rabbi Immaneul Jacobivits decry the use of scientific acumen in selecting sperm or performing genetic testing on the grounds that "Man..can never be the mere product of laboratory conditions and scientific ingenuity" (35).

The Protestant tradition is much more diverse than these two, having endorsed the controlled use of artificial assistance to fertilization (Lutheran) and condoned genetic engineering of somatic cells to correct genetic defects (cf World Council of Churches documents [36]). These three traditions coalesce in the form of various position statements on "procreative rights", most notably in the letter signed by 27 members of the clergy regarding genetic manipulation (37). While later renounced by at least one signatory (McCormick) this document underscored the abhorrence of representatives of the major faiths for human control of procreation as achieved by direct manipulations of the germline cells (egg and sperm).

For each of these traditions, the focus appears to be on the possibility that further knowledge of human genetics will inevitably lead to human distortion of a sacred plan, or to objectificaton of the process of procreation, a concern voiced 15 years ago by Leon Kass and before him by Jacques Ellul. Both thinkers reflected on the growing power of technology to redefine and engineer man, and the dearth of a "logos" or guiding principle to shape new decisions (cf 38). The claim that there is a right to an inheritance free from engineering is thus a second level concern about future applications of the genome data.

The major problem posed by biomedical advances is seen by major religious thinkers as the potential loss of human values stemming from a technologization of what were previously exclusively emotive, inchoate and fundamentally human processes. For Kass, the concern has focused on the likelihood that new technologies, particularly in genetics, would press parents and the obstetric population into a concern for "quality" in offspring and the transmogrification of procreation into a scientific abstraction of the laboratory. Kass observed that

> "Procreation is not simply an activity of the rational will. It is a more complete human activity precisely because it engages us bodily and spiritually, as well as rationally. Is there perhaps some wisdom in that mystery of nature which joins the pleasure of sex, the communication of love, and the desire for children in the very activity by which we continue the chain of human existence.?" Would not the laboratory production of human beings no longer be human procreation? (39).

Secular groups have taken similarly deontological positions on manipulating the human genome as well. In the instance of the European Human Rights Convention, the delegates of the Council of Europe meeting in Strasbourg, January 26, 1987 recommended that the Convention adopt the position that all persons have a "right to a genetic inheritance free from any form of engineering" (40).

Some philosophers, notably Joseph Fletcher take the opposing view by insisting that procreation becomes more human when it is more rational (41). According to this humanist tradition, reproductive choices are already in the realm of the conscious choice and to expand them, as would occur if more genetic data were fed into the decision-making, is simply an extension of human control.

The link between the earlier concerns of religious philosophers which was largely focussed on "artificial" reproductive technologies such as in vitro fertilization and the new genetic knowledge comes from the possible further erosion of human values by the passive movement of new data into the reproductive setting. In a traditional deontological sense, the precedent of allowing some genetic knowledge to be used (eg, in selective abortion) is seen to lead inevitably to the incorporation of ever more refined and trivial data in the reproductive process (cf 42). For Jeremy Rifkin, the issue is that the genome data will force parents to make calculated decisions about their offspring based on a so-called "genetic readout" (43).

The key issue for critiquing these deontological claims is to ascertain the likelihood that basic data from the human genome project will be used this way. As discussed in the section on the Context of Research, previous practices suggest that gene sequence data will be incorporated into reproductive decision-making. The ethical question is when such use is inappropriate, and if and when limits should be drawn.

c. Consequentialist Arguments

Control of embryogenesis, normal and deviant processes of development, direction of protein conformation and function, and other basic questions will potentially be resolvable through application of the knowledge generated by the human genome project. From a consequentialist viewpoint, the value of this knowledge of "basic biology" would be assessed in terms of its future utility. But, as many researchers have emphasized (cf 44, 45), the likely impacts of this fundamental knowledge can only be partially anticipated. A full analysis would consider the cost of its acquisition measured against the likelihood of benefit. However, since the amalgamated societal value stemming from the distribution of all advances from the genome project are unknowable, this analysis must be limited to the immediate prospects for genome data use.

In the first decade of discovery, it is highly likely that the knowledge base generated by this enterprise will create an industry centered on diagnostic devices which anticipate underlying illness, disease and incapacity. The first uses of this knowledge base, as previously discussed, have already occurred in the areas of prenatal diagnosis and genetic testing for the carrier states of major handicapping disorders.

The subsequent applications of gene data will be concentrated on broader diagnostics and therapeutics. Much of this next round of data will be of inestimable value to those in the private sector who need to anticipate medical costs, design insurance premiums and set rates for workmen's compensation, disability insurance and catastrophic health insurance. It is less likely that the benefits of such a program will accrue fully to those who are actually ill, since the technologies greatest power is in its ability to anticipate costly and presently only marginally treatable disease. Even here, the detection of individuals who test positive for a gene-based disease (especially if such a test precedes the onset of symptoms) may itself be a disvalue, especially when they (or society) are unprepared to deal with the data.

d. Conclusions

From a religious perspective, resolution of the fundamental deontological question of the acceptability of acquiring and using genomic information hinges on the degree to which manipulations of genetic information are violative of fundamental precepts concerning the separation of human and divine interventions. In a more secular sense, the deontological acceptability of acquiring knowledge of the genome turns on the question of whether such a process is fundamentally wrong, or violates basic human values. Since the acquisition can be made from cell lines or cultures, the gene sequence data need not involve violation of individual integrity or privacy. Similarly, the core material itself can be said to be evolutionarily neutral, and as such is not infused with intrinsic values. Thus, little objection to the research per se can be mounted from a deontological viewpoint.

The more general legitimacy of deontological concerns can be evaluated on the congency of the claim that use of genetic data in limited, more acceptable horizons (e.g., in identifying couples at risk for transmitting serious genetic disorders) will inevitably lead to more egregious applications (e.g., in identifying individuals for whom procreation will be proscribed by the state). The tendency for erosion or slippage of more serious from less serious moral error (eg, the prediction that liberalized abortion will lead to a diminution of respect for newborn life) can be checked against real-world events. In the present circumstance, there is little evidence that genome data will accelerate or cause moral disrespect for the unborn, or violate fundamental human values or human dignity.

Many issues are not readily resolved by appeal to either a deontological or a utilitarian calculus. Utilitarianism is limited where it relies on assessing a project purely on the basis of the favorable ratio between the net benefit of identified goods and possible harms. This is particularly so where serious harms might accrue to a few individuals but be justified by a large aggregate good. Similar limitations apply to a strictly deontological approach in which proponents insist that any duty to perform sequencing would depnd on showing that the thing in question (i.e., the genome) is a good in and of itself.

The relative weight to give the sometimes opposing positions derived from such perspectives is best resolved by examining the major ethical principles potentially involved with the genome project's objectives.

3. Major Principles of Ethics

Among the major principles operating in medical/social contexts in the United States are those of autonomy, beneficence and justice. While initially identified in the course of defining the major ethical precepts underlying the protection of subjects of human experimentation, genetic screening, and prison or fetal research, they have also found utility in discussing societal obligations for the provision of health care (46) and hence provide a suitable starting point for discussing the ethical implications of uncovering the human genome.

a. Autonomy

Autonomy refers to the capacity of competent adult persons to formulate life plans and decisions freely. In this sense, autonomy depends on being able to act deliberately, based on ones own judgment about the consequences of certain behaviors and their value or utility to oneself or others. Autonomy also can be regarded as a right (a form of liberty). Derivative principles intended to encourage or protect autonomy include privacy, confidentiality, protection of bodily integrity and a so-called "sphere of privacy" or personal space to which each person is presumed to be entitled. In medicine, respecting autonomy calls for allowing persons to participate actively in choosing a course of medical treatment for themselves, to elect termination of treatment, and to be given privacy and confidentiality. In sum, to be afforded respect as persons in the full sense.

The genome mapping/sequencing project directly impacts on the issue of autonomy in the areas of privacy and confidentiality where genetic data may be obtained which has immediate medical, social or familial value. Examples include the protection of an individual's genotype against unwarranted access or use, and the related question of permitting individual subsets of genetic information to remain sequestered (i.e. anonynous) or otherwise protected. A more sensitive issue pertains to the question of the ultimate sanctity and privacy of individually identifiable, racial or ethinic genotypes. The most compelling issues pertain to the national policies that determine when if at all to permit governments or private interests universal access to blocks of data pertaining to individual genotypes and how to protect the resulting data.

The primary allegiance to protecting privacy and individually identifiable data are especially complex with genetics. Often, it is possible to deduce an individual's likely genotype from that of family members. Similarly, it may be necessary to know a parental genotype in order to adequately counsel a child or sibling. Tensions arise where such information is of critical importance for others, such as where first degree relatives at risk for a serious but late-occurring disorder wish to know the genotypes of their nearest of kin.

Individuals may have the right to be treated autonomously, but can nonetheless lack some elements of the capacity to act autonomously. The genome project's data may impact on this capacity in several ways. The revelation of genetic data about an individual's future likelihood of disabling illness can change the way in which that individual is treated in society, much the same way as revelation of a positive AIDS test can now compromise individual liberty.

Genomic data can also reveal the presence of an incipient illness that could directly affect mental integrity and reasoning as in the instances of Huntington's and Alzheimer's disease. In these and similar circumstances, the existence of certain genes within an individual genome that predictably alter an individual's judgment and mental capacity may themselves disrupt autonomy. The existence of such "competence-altering genotypes" create novel ethical questions about the genome project.

By inference from animal studies, particularly the mouse, a predictably large universe of neurologically active genotypes exists in humans that could impact on fundamental elements of competence (47). The ability to anticipate individuals who have late-acting neurological mutations creates a new sphere of responsibility for protecting the confidentiality of genetic data. Any given competence-altering genotype will carry its own imperatives for use from the perspective of society, the individual, spouses, or employers, depending on the degree to which a likely impairment affects individual judgment or actions that impact on others.

When such data are available in advance of the manifestations of the disorder they can be used to counsel and assist individuals in making major decisions about their future lives and their reproduction. But when the data are accessible to others, the genetic information may be used to restrict opportunities (e.g., limited employment) and to diminish the standing and authority given to personal decisions made by the affected individual. For instance, early withdrawal of legal competence could be a foreseeable consequence of prescriptive testing for autonomy-affecting genotypes like those for Huntington's or Alzheimer's disease. Stigmatization and labelling are thus clear and present dangers posed by lax protection of individually identifiable genetic information.

These risks are most likely to be encountered in the case of genetic data which define behavioral disorders, and as such raise the issue of protection of the rights of the mentally handicapped. Particular care is necessary to monitor application of genetic testing to adults in families with affective disorders or other forms of mental imbalance or retardation. The counterbalancing claim of need to know arises when genetic data are sufficiently predictive and the nature of the illness constitutes a clear and present danger for others. In such circumstances, release of information may be justified.

For instance, were one of the genotypes known to predispose to seizure activity in mice be found to occur in an analogous form in humans, the disclosure of that data prior to issuing a driver's license (particularly for public transportation) might be warranted depending on the severity of the predicted disorder and the likelihood of its occurrence.

b. Beneficence

Beneficence operates primarily in care-giving settings in which practitioners are expected to adhere to Hippocratic maxims to act to benefit and to avoid harming their patients. The duty to act so as to benefit others is based on the premise that there is a positive duty to help others in furthering their own interests when we can do so with minimal risk to ourselves.

It is important to note that in many circumstances in which the acquisition or use of genetic information will occur (e.g., population genetic screening), a doctor-patient relationship may not apply. Similarly, in certain settings (e.g., occupational pre-employment testing) the rules for acquiring and protecting individually identifiable information may differ from those in a hospital or physician's office. In such circumstances, it is likely that special protections will be necessary to ensure individual benefit.

In the context of gene mapping and sequencing of population-wide data, scientific professionals might be expected to take steps to assure a maximal likelihood that benefits and not harms derive from their undertaking. In practice, this might mean safeguarding particularly sensitive genetic data (e.g., that indicating persons at risk for manic depressive psychosis in the Amish) or limiting promulgation of data with a high potential for misuse (e.g., genes unique to particular ethnic groups that might lend themselves to military application, or less egregiously, to employment discrimination).

The presupposition of beneficence is that the ends will in fact benefit an intended target population. A major issue in genomic testing is benefit to whom and by what measure? Since at the outset the research is not directed at identifiable persons or ends, initially, little realizable benefit will accrue. This means that the first donors of genetic material (cells or tissues) are extremely unlikely to reap any benefits from their participation. For this reason, it might be acceptable to use standardized cell lines for the first test systems. Such a move might not only protect individuals, but might increase the probability of accurate replication.

Beneficence exists along a continuum between goods and harms. Non-maleficence is the paired principle that refers to the duty to avoid harming. A ranking of objectives in applying this principle has been identified by William Frankena (48). Frankena believes that there is a greater obligation 1) to avoid harming than 2) to prevent future harms; and 3) a greater obligation to remove or mitigate existing harms than 4) to promote good. From this typology, it is more important to identify and redress the consequences of extant harms than to prevent their occurrence in the future.

For the genome project specifically, Frankena's ranking of duties would imply an obligation to ensure that the acquisition of genetic data did no harm, since the likelihood of deriving immediate good from such research information is extremely low. In the context of applied genetics, this hierarchal treatment could be taken to mean that there is a greater duty to treat individuals with serious genetic disease that to prevent their birth, and in turn that it was more important to prevent the birth of a child with a severe genetic handicap than to limit procreation of carriers of the same gene in the future. Frankena might also argue that it was more valuable to treat genetic diseases than to design programs to enhance human well-being through genetic engineering. Each of these claims, however, depends on the definition of "harm", since for many an abortion is the ultimate harm, while treatment (especially experimental therapy) can by definition be harmful.

A critical adjunct to Frankena's hierarchy is the demonstration that a given policy is the least harmful (i.e., least restrictive) option available to achieve a desired outcome. By this analysis, it could be argued that it is more important to have a genomic sequence base for all newborns as a reference to identify and counsel individuals at risk for handicapping illness than to identify them as adults when they have already embarked on life plans. Here, the risk of severe psychological problems would have to be anticipated, since newborn screening has been shown to produce severe anxiety among families in which a potentially handicapping condition has been identified (49).

From the context of applied genetics, Frankena's ranking would assign greater weight to a program to avoid the conception of a predictably damaged fetus than to design a program to ensure that all carriers of a given deleterious gene avoid reproducing. The latter policy would be given primacy because avoiding the birth of a single child is less restrictive than compelling all couples at risk for disseminating a deleterious gene from reproducing.

c. Justice

Justice is a complex and multidimensional concept which embraces theories of equity, distributional and retributional systems of compensation, and doctrines of fairness each of which have potential utility as applied to biomedical settings. At the onset of the genome project, the major issue of justice is that raised by the proportionally large investment in one phase of biological research, and the concomitant question of the allocation of initially scarce resources.

The major entry level justice issue is whether or not the substantial investments in government dollars anticipated for the discovery phase of the project (when applications are minimal) unduly penalizes or deprives a deserving group of needed health services or more directly tangible benefits. This problem cannot fully be resolved because of the fundamentally unknowable nature of the final distribution pattern of benefits. However, it is unlikely that any one group would be deprived of eventual benefits (as might happen if a disease-directed program were initiated, say for the Caucasian dominated disorders of PKU or cystic fibrosis) because of the universal applicability of the accumulated data set.

Genomic mapping may be useful in satisfying John Rawl's "second principle" of justice which calls for society to provide benefits to the least advantaged. Certainly, in a simplistic and minimalist sense, key genetic data will prove useful in assisting in identifying such persons. The subsequent, important question of how to perform that identification and how to protect those so identified, are subsidiary questions that would need to be resolved.

The basic premise of justice is that persons will be rewarded, protected or compensated in a way that is proportional to what they deserve, according to need, to social desert or merit, or ability. The core concept of equity presupposes that persons with similar dispositions or makeups be treated similarly. In a genetic context, this position would favor policies that tend to recognize and treat individuals with similar genetic compositions alike. In practice, this simplistic application of equity could be compromised by the complexities of genetic determination of important characteristics or disorders.

Pleiotropism in which a single gene produces protean effects and manifestations demonstrates that for all but the most severe gene-directed disorders, the actual end-points of similar genes in terms of individual characteristics will almost certainly vary to some substantial degree. The classic example of neurofibromatosis (von Recklinghausen) in which individuals may express the gene in a single cafe au lait spot through to multiple tumors and bony protuberances points up this dilemma.

In large part, the problem of applying principles of justice to genetic issues is intensified by this often great disparity between a genotype and its predicted phenotypic expression. The dependency of many polymorphic traits for interaction with other genes (e.g., diabetes, some forms of cardiovascular disease, hypertension, etc) also confounds programs that are intended to identify or compensate "victims" of atypical genes. In the past, this observation has meant that the gene has rarely been the unit of significance for policy formulation. A notable exception would be those policies which identify a genetically linked phenotypic characteristic, e.g., skin color and direct policies to assign social goods accordingly as in the Republic of South Africa.

In the context of genomic testing in its discovery phase, appeals to justice might be made on the basis of fairness in distribution of resources for probing particular genotypes, e.g., no one ethnic group's genomic features would be given undue weight for investigation, protection or other benefits. An argument for singling out a particular group could be made on the basis of compensatory justice. Such arguments might be used for providing enhanced attention to damaged or high-risk groups, in spite of the possible lack of generalizability of the resulting data. For instance, offspring of Atomic Bomb Casualty Commission identified survivors or victim's of Chernobyl might be given special treatment in the sequencing project to ascertain radiation induced deviations from the norm of human gene sequences.

Justice would also apply to the question of primary distribution of health resources. For instance, the genome sequencing project's budget might be designed to ensure that outlays not diminish the resources available for more directly health beneficial activities related to genetic disease. Were the project's funds to compete directly with other outlays in genetics, an arugment of proportionality might be invoked to support the claim that the project's expenditures not be at the expense of continuing support for a previously identfied group that had received funding. The more difficult question of the distribution of benefits between basic and applied research would also need to be addressed. Justice would require that the genome project's outlays not be at the expense of compensation or screening programs already in place for extant populations with serious handicapping conditions attributable to genetic disease.

These examples ask for proportionality for expenditures earmarked for the genome project and those already budgeted for parallel programs in the genetic sciences, on the basis that comparable conditions require comparable support. Here comparability is limited to the gross designation of a genetic basis for the condition(-s) in question. But for setting priorities among programs which are not alike in terms of their genetic basis, no argumentation can be offered that effectively claims that outlays for the genome project should be at the expense of more distant health-related expenditures. On this basis, every program in the basic sciences would lose out to the much more pressing needs of relieving human suffering from disease.

A secondary issue of justice concerns how the government investment in research funding is to proceed once it is decided to proceed with the project. Simple equity would argue for a fair distribution of grant money proportional to the demonstrated proficiency of the selected laboratories. In this area, the issues are not novel nor differ to any substantial degree from those inherent in funding any other government funded programs except that the project will have immediate utility and universal value to many other researchers. In this sense, the major ethical issue that exists in the genome project relates to the fairness of the program's ultimate distribution of benefits. It is here that the balance between private and public "rights" to knowledge comes into play.

The claim for public access to personally identifiable information is a traditional problem in medical confidentiality. It assumes the dimensions of an issue in justice when the value of that information extends to groups or society as a whole, for instance as when the disclosure of AIDS in a surgeon or other individual with wide public contacts appears to impact on the community. How should the discovery of a particular constellation of genes that identifies a potentially handicapped individual or population be handled?

A genetically-based system that incorporated principles of justice could be devised. For instance, a policy to apply a gene screen at birth to determine those individuals in greatest need of remedial services in their immediate prepubertal years would recognize previously mentioned second principle of Rawlsian justice that called for adjusting policies in favor of the least well off. Such a program already exists in the form of PKU screening programs. The rules and requirements for fair and equitable application of such a program might require (as was the case for PKU), that its cost-effectivenss be demonstrated, and that the operation of the program did not compromise or impair the ability of the health care system to provide services to persons with less severe handicaps.

Until now, American values have operated in the guise of "gene-neutral" policies, regarding classifications based on gene-determined features either as inherently suspect or to be counterpoised with policies of fairness, as with affirmative action programs by gender. While very few policies overtly identify a genetic condition as qualifying (or disqualifying) an individual otherwise deserving of receipt of protection or services, the existence of a gene bank of readily identifiable characteristics may invite a wholesale reformulation of protective policies such as those ensconced in the Equal Employment Opportunity Act, the Civil Rights Act, etc.

To be in accord with major principles of justice, such policies might have to accomodate the genetically handicapped preferentially, as was done in the Rehabilitation Act of 1973. Even under the protection of such laws, the possibility still exists for discrminatory use of genomic data since the laws permit bona fide qualifications that may have a genetic base to justify non-employment (50).

This eventuality could arise whenever a genetic marker provides a seemingly biological basis for policy that avoids the legal pitfall of "invidious discrimination". Programs which directly acknowledge genetic differences will thus not be seen to be wrong on their face, as long as they worked to compensate for genetic disadvantage. Were they to work to limit the rights or opportunities of genetically affected individuals, their operation would be ethically suspect.

Justice could thus be taken to dictate that practitioners ensure that each person be given maximal opportunity to learn his genetic heritage. The relevance to the genomic project is clear, but the question of when genetic information reaches a threshold of critical value such that it is disseminated even to those who do not request it will be one of the earliest ethical issues of the application phase of the project.

The simple dissemination of genetic knowledge as is currently done via genetic counseling presupposes a continuing policy of disclosure only upon request. Such a policy also presumes the neutrality of the counselor-client interaction, in which little or no coercion is exerted by the information-giver. While most clients report that they were not influenced by the counseling encounter in their decision-making (51), the vulnerability of genetically at-risk couples to manipulation remains a major issue. The client centered approach of most genetic counselors (52) limits any major systematic influence on the outcome of procreative acts by deleterious gene carriers. This value orientation would be severely strained by the advent of vastly more relevant data indicating which procreative decisions were likely to have the most damaging--and those least damaging--effects on the gene pool.

The appropriateness of any policy directed at instructing genetic counselors to take account of the "genetic load" of mutations, and the responsibility borne by those carrying such genes, depends on the moral weight to be assigned to the intergenerational transmission of mutations.

d. Intergenerational Responsibility

A major theme of ethical analysis is the question of intergenerationalresponsibility. This issue will become relevant as the genome project generates data which permit predictions about the consequences of present acts on the future frequencies of various genes. To control the future requires control over the propagation of identified genotypes. By necessity those genotypes will be in persons (at least initially). Hence, to control future genotypes will entail constraining or encouraging selective mating or conception. Any such scheme will entail ineivtable questions about coercion and reproductive freedom.

The priority to give these concerns is presently low, principally because the claims of already extant persons (e.g., for reproductive freedom) are constitutionally protected, at least currently. Many ethicists also believe that obligations to individuals currently alive takes precedence over those to future persons. In keeping with the priority sequence developed by Frankena , policy persons may rightly insist on the primacy of not-harming in present generations over the obligation to prevent future harms.

But the genome project may well generate data that point to serious future consequences to society of certain mating patterns. In such circumstances, the tensions between autonomy and justice will have to be resolved. It is now clear that alternative systems of procreative control will have divergent impacts on the gene pool. A program which allowed assortative mating to proceed without use of genetic data that could provide individuals data on carrier status of potentially deleterious genes would produce more mortality (and medical costs) than would one in which carriers were alerted to their status and encouraged to mate with non-carriers (19). A less extreme policy, analogous to programs already in place for Tay Sachs disease testing, would be to inform carriers only when married to each other and at risk for having an affected fetus.

The net effect of this latter policy when coupled with reproductive compensation to make up for the loss of fetuses homozygous for the recessive disorder in question, is to increase the genetic load. Ironically, the most extreme "eugenic" policy which denied carriers the privelege to mate would sequester or mask the most deleterious genes over several generations of operation. Over time, the consequences would be an increase in both the frequency of deleterious genes and (should less restrictive policies be re-imposed), a great increase in the number of deaths due to homozygous recessive disorders. Hence the most autonomy-respecting policy might be the most costly, while a more restrictive one which appeared to be the most efficacious in the short run, the most damaging over time. Only the most restrictive policy, for instance one that denied procreative opportunities or counselled chastity indefinitely, would have a truly eugenic effect (19).

Such eugenic applications of the gene sequence data must be taken seriously, since how policy makers decide to use genomic data to identify carriers of "defective" genes will potentially affect their contribtution to the gene pool. For instance, if carriers are required to submit their genetic data to insurance companies underwriting health or disability benefits, indirect restrictions on child-bearing could result from requirements to avail themselves of amniocentesis, to pay a premium for full benefits, or to subject themselves to genetic counseling as a condition of receiving coverage.

Several anticipated applications of gene probes will in fact identify and permit the early remediation of conditions which are extremely costly to treat and are transmitted as dominant disorders, (e.g., retinoblastoma; Wilm's tumor; familial polyposis coli). In these instances, some increase in the frequency of the disorder in future generations is likely. Except for the procreation of individuals with dominantly inherited or X-linked conditions that are presently maintained only through reintroduction of the gene by mutation each generation, e.g., Duchenne muscular dystrophy, (collectively a very small number of persons), the possibility of serious erosion of the gene pool through misuse of genome data is presently not of eugenic proportions.

4. Major Arguments for Expanding Existing Mapping/Sequencing Programs

a. Issues of Scale

From an ethical standpoint, the key arguments for scale turn on the distributional costs and benefits of the project. If the public is to be asked to pay an appreciable portion of the cost, then arguendo, it would deserve to participate in the priority setting and to ask the scientific community to justify the decision to embark on an expensive, full scale project. If the largest share of the costs of the project are to be borne by industry, then few if any questions of priority will be posed other than those chosen by the persons investing risk capitol in the project.

The ethical issues of the dimensions of the mapping project are also shaped to an appreciable degree by the end uses to which the project is intended. The implications of a broad mapping project, in which the major genomic regions are simply charted with little or no resolution of any but major disease-conferring genes versus a refined one in which the full sequence is revealed hinge on the relative value to be assigned to directed versus non-directed (basic) research.

Thus, while it is likely that even a five kilobase map would have great utility in permitting the identification of markers linked with major diseases, and might thereby provide the kind of detail of interest to insurers or others bent on identifying with certitude "high-risk" individuals, it would have appreciably less scientific value to basic researchers. Public interest groups have pointed out that the project has little immediate return, andwould be extraoridnarily wasteful at a time of increasing economic downturns.

Leroy Hood and his co-workers at the California Institute of Technology have observed that sequencing the entire genome could be regarded as a wasteful and trivial exercise (53). This interpretation is predicated on the estimate that 95 percent of the DNA in which the approximately 100,000 human genes is presumed to be sequestered is "junk" DNA with no known functional significance. Molecular biologist Robert Weinberg has reinforced this viewpoint by asserting that most of the information that will be gleaned from a full-scale sequencing project will be meaningless (54). In his view, "much of this [genetic] material seems to represent evolutionary detritus, dead ends, side alleys and parasitic sequences that are extrinsic to the functioning of our genes."

This argument is narrow and focussed solely on the functional significance of genes. What is junk and debris to one scientist is invaluable data to another. In the broadest possible view, the genome project's information base will be a treasure trove of data for those in the evolutionary sciences and the field of human genetics generally, as the very "extraneous" material described by Weinberg is found to provide clues to human origins and the overall construction plan of the human genome itself. Smith and Hood have also countered the type of critique offered by Weinberg by pointing out that no general procedures can distinguish the two (or more) forms of DNA.

b. Arguments for High Resolution Mapping:

The proponents of making a major effort to map and sequence the human genome argue on grounds of technical readiness, intrinsic values for medical science, the good of knowledge in and of itself, and the large potential for secondary applications of the resulting knowledge base. Smith and Hood point out that it would also be more efficient to pursue the full map in lieu of a selective project.

c. Technical readiness:

From an ethical viewpoint, the issue of readiness implies elements beyond simple efficiency and cost-effectiveness. From this perspective, readiness implies that there has been adequate assessment of secondary factors, e.g., the degree of social and cultural preparedness for potential impacts of the project; possible skewing of societal priorities; and the relative value to be assigned the project in the context of other projects. But it is in the nature of basic research that all of the seondary impact cannot possibly be foreseen. Hence, the claim that there are "unknown" but potentially harmful consequences cannot be given a quantifiable weight.

d. Values for medical science:

The strongest scientific argument for proceeding with a reduced-scale project is that by providing only broad, linkage group detail through mapping and RFLP analysis, sufficient data will be provided to do most of the important tasks identified as medically important (e.g., identify genetic diseases, markers and receptor sites). This viewpoint fails to recognize that the gene sequence would provide a powerful tool to scientists bent on elucidating the fundamental relationship between structure and function and therefore would have substantial value for medical science over the long-run.

e. Intrinsic values

These arguments are typically deontological in that they appeal to the good in and of itself of acquiring knowledge about the human body. The strongest argument for pursuing the human genome project is thus that it will render knowable much of what determines the human condition. In the words of Lloyd Smith and Leroy Hood, "sequencing the human genome would provide one of the most powerful tools mankind has ever had for deciphering the mysteries of his own existence" (53). But others are not nearly as sanguine. Robert Weinberg doubts that the genome project will unveil much about the mysteries of life. According to Weinberg, "It is fair to say that most of the stumbling blocks currently in our way in the difficult problems of human biology will not be pushed aside by the sequencing juggernaut" (54). While a genome sequence is unlikely to be a sufficient description of basic biological mechanisms, many philosophers of science would argue that at some point it is a necessary component of any full-scale analysis of the integration of the functioning of the organism.

f. Counterarguments

This proposition itself is philosophically questionable, since it presumes that much if not most of higher human attributes are "reducible" to molecular constructs. This issue is in serious doubt. Many researchers now question whether higher human functions, including neurophysiology and psychology, can be reduced to (i.e. derived from) prior knowledge of molecular constructs. This issue is particularly relevant to the question of what will be "revealed" by a fully elucidated human genome. A reasonable hypothesis is that while little of direct or immediate value will be obtained about complex phenomena, insights into the possible construction of control regions for embryogenesis (e.g., the so-called "homeo-box"), the genetic basis for organizing neuronal pathways, and the genetic control of sexual differentiation will all be signficantly enhanced by a full-scale knowledge of the genome.

g. Spinoffs

Such a tool would be invaluable for answering fundamental questions about biology such as the genetic basis for neuronal function and distribution, the basis for "self-recognition" in immunology, and the biological roots of behavior. This type of argument can be termed the "spin-off" argument. In this type of reasoning, no one can assign an upper limit to the value of the research because no one can anticipate fully the range of secondary values to the project. For instance, researchers at the Lawrence Hall of Science at Berkelely, California postulate that "the mapping project will produce as many spin-off projects as the space project has" (55).

The relative value to be assigned to "unforeseeable results" is difficult to assess. Philosophers of science as well as practitioners like John Platt have emphasized the intrinsically greater value of hypothesis formulation and testing over blind assembly of unrelated datum points, material he calls "bricks in the brickyard" (56). The human genome project appears much more like the latter class of research, and hence is criticized on the ground that it is so much "shooting in the dark". However the efficiencies of scale coupled with the opportunity costs of not pursuing a full description, and the risk that some or all of the data may be acquired by foreign or highly proprietary interests, argue for embarking on a full scale project. The strongest argument in the end is that the basic knowledge assembled by the project is a necessary concomitant to understanding human origins in an evolutionary sense, and the "grand design" of our genetic heritage. The undoubted existence of "meaningless" genetic data and redundancies itself carries information about the assembly and operation of an integrated and functional genome. Without all the pieces, we will lack the data needed to assemble the whole picture.

A more cautionary view would hold that because we will inevitably uncover genetic data which is both unwanted and subject to nefarious use, we should proscribe a full-scale search until we have charted the domains where the most dangerous gene sequences are likely to lie and instituted appropriate safeguards or blocked their acquisition altogether. This argument presupposes that we can know in advance what information is "intrinsically dangerous". It also presupposes that if we could chart the locale of a gene that was linked, with say violent, antisocial behavior that that gene would in fact be predictive in and by itself of a the higher level phenomenon of concern. These possible resolution of these issues will be discussed in the following section under "Level 3 Considerations".

5. Long-term Implications

If a full-scale project were to be undertaken, its long-range impacts and consequences for future generations would need to be considered as part of an ethical analysis. Phillsopher-ethicists like Daniel Callahan, Director of the Hastings Center have argued that scientists bear a moral responsibility for using what he terms "moral imagination" to anticipate the full ambit of uses and consequences of their work, especially when that work is in the basic sciences (57). The human genome project's implications are likely to be vast, and impacts in the public health, insurance, and medical diagnostic areas we have already discussed are properly among those that would need to be anticipated prior to its implementation.

Key to any such analysis is a careful evaluation of what we would know if we were successful in dissecting the human genome. While some of this knowledge is by definition unknowable, we can chart in a broad way where we will be after making the sequence a coherent whole.

In this sense, it matters significantly if the major genes that will be found simply expand our knowledge base of human genetic disease (level 1); if they can be used to identify and act on negative human disease or behavioral proclivities, attributes or susceptibility states (level 2); or if the uncovered genes can be used to assemble a picture of human traits, abilities and higher order functions implicit in personality, criminality, intelligence and aptitude (level 3).

a. Level 1 Considerations

Most of those commenting on the genome project believe it is primarily a project which will uncover the basis for human gene-based disease. It can be argued that this minimum expectation provides a necessary but not sufficient rationale for a full scale project since most of the major disease-conferring gene sequences are discernable within chromosomal fragments or RFLP-marker delineated regions that have already been defined. In this sense, revealing the the full ambit of genes between regions or areas of concern is conceivably a substantial waste of time and effort.

Until recently, most genetic disorders were diagnosed through physical or biochemical-genetic examinations of probands, detailed family histories, and clinical laboratory tests including both cytogenetics and biomistry. The movement to diagnose patients based on analysis of DNA samples from patients and their families is now increasing in popularity. This new development represents a remarkable advance with drmatic implicaions for the genome project at the level of diagnosis of human disease. Specificaly, a DNA-based test permits pre-clinical detection of disease, prenatal diagnosis, determination of carrier status, and precise delineation of the condition in question from disorders that might mimic its outward manifestations (phenocopies).

The extent of applicability of this new development once the full human genome is deciphered is limited only by the philosopical question of the relationship between defects at the gene level and human disease. It appears likely that a significant fraction of the major single-gene diseases, particularly those with a metabolic basis will lend themslelves to this analysis. The most recent development is one in which a short sequence of a patient's DNA can be greatly magnified to permit hybridization studies with a specific probe. The immense expansion of the number of available probes will likely be one of the most signfciant results of the genome project .

The signficiance of these observations is that to the extent that the human genome project rapidly increases the number and availability of DNA probes, the detection, prevention and ideally treatment of human genetic disease will be comparably enhanced. The most important technical question is the necessity of deciphering the entire genome just to identify the sequences of interest and concern for diagnosing or preventing this category of disease. In this sense, the resource question previously raised in the discussion of Justice is pertinent, since the available resources for public health work generally might well be limited by the sequestration or concentration of effort in the genome progject, at the expense of doing more directed research aimed at identifying the major environmental and/or interactional factors that shape a particular health problem (e.g., lead poisoning).

As a broad guideline for resolution of this tension, it can be argued that policy-making bodies have a primary duty to compensate or provide retribution to individuals who have already been harmed through the misuse of genetic data (either by neglect or overt acts of discrimination) before they were to embark on a full-scale project in basic genetics. Secondly, implementing agencies (e.g, the NIH or DOE) could be required to provide "turn-outs" from the full-scale project so that intrinsically valuable data, e.g., that related to the cause of a serious human disease, could be rapidly developed prior to the completion of the entire sequence. Third, Congress could be asked to assure that funding be taken from "new" monies and not from existing programs intended to provide more direct health benefits to the public.

b. Level 2 Considerations

If the genome bears regions and gene sequences that are intimately associated with major, complex human disorders such as heart disease, hypertension, and mental diseases like schizophrenia, then ferreting out the full sequence is more justified. In this sense, the genome project can be envisioned as an effort to partition out the complex genetic components of multifactorial diseases, including the basis for susceptibility to them. This work raises the previously identified issues of privacy and confidentiality and highlights the dilemma of having the most revealing data sequestered in the private sector. It is conceivable that a regulatory body would be assigned the responsibility for monitoring all Level 2 applications of the genome project to protect against potential abuses such as those associated with sale and use of workplace screening probes.

Implementation of this policy could take advantage of broad existing mandates, such as the Pure Food and Drug Act and its subsequent forms that regulate genetically engineered products such as Human Growth Hormone and gene-directed medical devices such as diagnostic agents based on DNA probes. In theory, a particularly thorny product such as a probe directed at detecting an ethnically based predilection to workplace hazards could be tightly licensed and regulated much as radioactive contrast agents are currently controlled.

c. Level 3 Considerations

Recent data from twin studies carries the inference that complex behaviors and attributes including activity level, alcoholism, anxiety, criminality, dominance, estraversion, intelligence, locus of control (i.e. personal autonomy), manic depressive psychosis, political attitudes, schizophrenia, personal values, and even vocational illness are under significant albeit indirect genetic control (59). If true, these observations suggest that a full study of the human genome could unveil gene clusters or combinations that are associated with human predispositions and value-orientations of extreme sensitivity and political weight. Under some circumstances, ethicists have argued that data like these which have great potential for nefarious use should be considered "off-limits" (cf 57). Examples would include the development of indiscriminate and/or highly contagious biological warfare weapons, or the study of pharmacological techniques to make individuals vulnerable to psychological suasion. But even with the most nefarious uses, others can envision potential benefits of research, such as spinoffs of biological warfare research in the development of vaccines, or the development of better understanding of human will and autonomy.

A position to limit research in the genome project has by necessity had to appeal to an apocalyptic view of the uses of genetic science. Such a viewpoint is evident in the writings of Jeremy Rifkin and has previously been espoused by Herman Kahn, the now deceased Director of the Hudson Institute. According to Kahn "genetic engineering has in it the makings of a totalitarianism the like of which the world has never seen" (59). To prevent some or all of this data from getting into the hands of potential despots, Kahn proposed creating an "index of forbidden knowledge." Presumably Kahn was concerned about setting certain limits on what gene-based associations would be gathered and encoded into personally identifiable documents or data sets. Whether or not such potentially powerful data would be grounds for banning its acquisition has been discussed previously in the context of the recombinant DNA debate.

Some individuals had argued for an early ban on rDNA research on the grounds that serious and potentially disastrous and unforeseeable consequences could result from its activities (60). Philosophers like Carl Cohen have argued that the only conditions under which a ban on acquistion of knowledge could be philosophically justified would be when and if such knowledge could reasonably be determined to pose a clear and present danger to identifiable persons or groups.

Objections to initiating the research would be effectively countered by demonstrating in advance of such a project how such worse case scenarios would be countered. In the case of DNA research, the claim was that simple prudence called for demonstrating that sufficient precautions would be put in place to protect against untoward events prior to initiating the work. Such precautions (e.g., biosafety committees, environmental release guidelines, etc.) are singularly inappropriate for genomic research. Here the problems are more likely to be generated by the applications of the data, and safeguards to protect ethically sensitive issues would be both necessary and sufficient. It is possible but unlikely that the resulting precautions could be so costly or extraordinary that the project itself would have to be abandoned or reconsidered purely because of the scale and additional cost of imposing protections.

As a general rule, it would be reasonable to impose safeguards that were proportional to the level of likely applications of the resulting information. Thus, if the data uncovered by the genome sequencing initially applied primarily to human disease, then the traditional safeguards implicit in the doctor/patient relationship and genetic screening generally would conceivably suffice. In this instance, the guidelines developed by various groups for protecting subjects of mass genetic screening programs would be pertinent.

The first set of recommendations was promulgated by the Genetics Research Group of the Hastings Center (then the Institute of Society, Ethics and Life Sciences). These guidelines specified the safeguards of privacy, confidentiality and community protections implicit in a large-scale screening operation (16, cf 62).

The second set of recommendations emanated from a committee of the National Academy of Sciences. These guidelines emphasized the traditional protections of human subjects and argued that protections on confidentiality and privacy of the resulting data were sufficient governmental limits on the applications of genetic knowledge (62).

One key pronouncement of the Hastings document was that genetic data not be used as a basis for defining normalcy in deservedness for social programs or services. This statement was predicated on the premise that equal standing and opportunity were extrinsic to genetic or other biological data that indicated an individual's race, genotype or ethnic affiliation. Provisos of this sort may have to be re-examined in light of the findings from a genome project which are likely to show significant associations between genetic makeup and human functioning. Of course, the fundamental deservedness of individual's for social support or programs should be unaffected by any butthe most relevant gene-based aspects of their makeup, e.g., the qualification of a child with Down syndrome for placement in a home for the impaired. .

The extension of genetic information into the functional aspects of an individual and society is predicated on that data carrying significant predictive information. This would hold true were some uncovered genes to follow the pattern established already for certain medical disorders. Thus it is likely that the project will uncover key gene sequences which may demarcate individuals at increased risk of catastrophic illness (eg, aortic rupture from Marfan's syndrome; focal-motor epilepsy; hemolysis; sudden infant death syndrome (SIDS); or the syndrome of unexplained nocturnal paroxysm and death in certain Asian populations).

In each instance the gravity of the risks to individuals or others may be sufficiently severe as to suggest restrictive policies for instituting testing (eg, compulsory test requirements); from another ethical viewpoint, these same conditions may warrant special guidelines for discolosure or protections of the most senstive data.

These issues become most compelling when specific, identifiable gene sequences known to have harmful consequences (e.g., those for Huntington chorea) are identified in single individuals. In such circumstances, the ethics of non-disclosure are problematic, since significant others (i.e.' members of the individual's extended or immediate family) are at risk of a potentially life-threatening illness. The justification for breaching confidentiality is greatest where the disorder is potentially treatable (unlike Huntington's at present), and the affected individuals are minors. Specific examples are discussed in the following Section on Eugenics, under "Privacy".

6. Eugenic Implications

The logical consequence of mastery and control over human DNA will be to raise anew the highly charged issue of eugenic applications. The major difference between the current enterprise and previous ones is that for the first time any putative eugenicist will be able to claim to be using "hard" genetic data to effect desired ends. With even modest first steps in the genome mapping project already achieved, genetic knowledge has already reached the level of sophistication to permit an active albeit limited eugenic program in animal husbandry. Most importantly, in animal husbandry, this achievement is dependent on the ready accessibility of embryo transfer methods, a modality that is fast reaching widespread public acceptability.

When Francis Galton defined "eugenics" in 1883 as "the science of improving the stock," he intended the concept to extend to any techniques which might serve to increase the representation of those with "good genes" (64). Thus he indicated that eugenics was "by no means confined to questions of judicious mating, but which, especially in the case of man, takes cognisance of all the influences that tend, in however remote a degree, to give the more suitable races or strains of blood a better chance of prevailing speedily over the less suitable than they otherwise would have had."

A major entry level question is whether or not any kind of significant eugenic effect could be achieved by use of the knowledge gleaned from the human genome project. At least one commentator, Rene Dubos declared in 1978 that "there is hardly any chance that the genetic constitution and the fundamental behavioral patterns of the human species can be signficantly altered by scientific interventions in the near future" (65).

The countervailing attitude was exemplified by Julian Huxley who identified three factors which would have to occur concurrently to achieve true eugenic impacts: at the present moment [it is] very difficult to envisage methods for putting even a limited constructive program of eugenics into efffect...this [is] due as much to difficulties in our present socioeconomic organisation as to our ignorance of human heredity, and most of all to the absence of a eugenic sense in the public at large (66)."

The advent of a full gene sequence, coupled with the expansion of reproductive technologies, provides a perhaps unanticipated avenue for direct genetic impact. While that impact is likely at first to be limited to individuals, it is entirely probable that utilization of gene sequence data will be sought under unusual circumstances not now foreseen. For instance, in the event of nuclear war in which substantial numbers of individuals had their germ plasm damaged by radiation, it would be plausible to use gene sequencing techniques to test the integrity of donor sperm and to select individual procreators on the basis of genetic tests.

Whether such use of the genome data in less extreme circumstances would be appropriate or ethically acceptable turns on the particular religious perspective which one holds.

a. Judeo-Christian traditions

From the Judeo-Christian perspective, human life is itself sacrosanct and a creative act of a divine force. Interventions to alter procreative outcomes, according to some in this tradition, constitute a violation of the injunction that only God created life. In contrast, interventions to repair, reform, or otherwise correct a seriously damaged child or person are considered eminently human acts. Thus, correcting a genetic defect (via gene therapy) is afforded entirely different standing and approval than is either altering the germ line of parents to avoid the birth of such a child in the future, or the improvement of genetically based features (e.g., cosmetic surgery to enhance the contours of the face).

Ethically, the focus of the application of genetic testing would appear to matter significantly. Were it to be directed at the procreation of couples with the end of identifying a few high risk individuals, it would be more consistent with theologically accepted practices than were it directed at screening gametes or fetuses in utero. The Judaic tradition seems to most strongly endorse the desirability of securing a eugenic effect through marriage and procreative choices. The Jewish tradition, "quite consciously strives for the promotion of the quality as well as quantity of the progeny" (67) For instance, the rabbinical tradition appears to condone the use of both moral and physical tests of acceptability of spouses as the basis of a choice of a marriage partner, as long as such testing did not disturb or contravene the relationship after its consummation. Genetic testing for Tay Sachs disease is a contemporary form of this practice. In Judaic tradition more generally, it was also considered desirable to anticipate (genetically) ones children, for instance by examining the maternal uncle of a prospective bride (68). The Talmudic injunction against marrying into a family of lepers, epileptics or related diseases has been interpreted to be an early recognition of eugenic principles (69).

The Christian tradition appears to recognize the importance of genetics in the prohibition against marriage of close relatives (70). The major citation to this effect is from the letter of Pope Gregory I to Augustine of Canterbury (601 AD).

b. Positive Eugenics

This definition carries a racial perjorative which more contemporary concepts of eugenics have tried to avoid. Most commentators have nonetheless rejected or cast heavy doubt on any uses of genetic engineering to "enhance" or directly improve the human condition. Thus, the President's Commission has declared that efforts to improve or "enhance" normal people (as opposed to ameliorating the deleterious effects of genes) is at the least "problematic"

(71). Similar questions apply of course to use of non-gene directed interventions, such as the use of recombinant DNA human growth hormone to augment the growth of short stature pre-adolescents, or the muscle mass of Olympic athletes.

It may well be that this problem has more to do with means than with ends, since the general objective of "improving" the human condition is roundly applauded by most if not all ethicists. Most concerns about eugenic policies in the past have focussed on the methods used to attain them (sterilization, Lebensborn breeding farms, etc), rather than critiquing the ends themselves.

In the past, philosophers beginning wth Plato realized that eugenic ends could be achieved through subtle or direct incentives to bring together presumptively "fit" genotypes. Positive eugenic effects (herein defined more narrowly as the achievement of systematic or planned genetic changes in individuals or their offspring which improve overall phenotypic adaptedness) could be achieved by programs that did not require direct manipulation of the genetic material.

In Singapore, positive eugenics is "practiced" by public policy (72). The selective mating of individuals with presumptively valuable characteristics such as intelligence, character, etc. is actively encouraged and the national policy regarding abortion and family planning is being watched to ensure that positive eugenic effects are realized. Persons who have limited education and who are willing to elect sterilization were identified in a 1984 law as deserving compensation ($4,000 U.S.). While it is true that mates with similar IQ's tend to find each other, there is no assurance (in fact quite the opposite) that offpsring will represent higher mean values: in practice, regression to the mean has doomed virtually all non-selective breeding programs whether in humans or animals.

Limited human "eugenics" could be practiced with less than extreme alterations in public policy. Useful analogies from animal husbandry suggest that even limited genetic knowledge coupled with control of the procreative activities of specifically identified individual animals can have dramatic eugenic effects. A recent study compared the relative value of directly engineering stockyard animals through germ-line engineering versus specific breeding patterns based on RFLP-identified gene segments (73). The author successfully showed that dramatic success in increasing the representation of "quality" genes in subsequent generations could be obtained throgh judicious mating of selected carriers. In view of the present widely held belief that germ-line engineering should be proscribed as a means of eugenic improvement, it is useful to review just how such a program might work beginning with the new data from the genome project coupled with even modest expansion of sperm banking activities.

The study alluded to began with the premise that traits of economic importance in animals are almost entirely polygenic in origin. Before the advent of defined genes and their molecular markers, this truth limited animal husbandry to crude measurements of phenotypic characters and their propagation by selecting animals presumed to carry the necessary genetic material to express them in offspring. These traits are keyed by so-called "quantitative trait loci", are likely to be widely dispersed throughout the genome on different linkage groups. Only with the advent of a reasonably complete map and sequence of the entire genome is it likely that these traits will be located. Once found, these markers can be identified in a given individual and suitable gametes (usually sperm) selected for fertilization. Back-crossing, in which the F1 generation is mated repeatedly to the paternal donor, increases homozygosity for these markers and increases the likelihood of the desired eugenic effect.

Given that the major traits considered by some commentators to be of greatest human value--intelligence, general strength and coordination, and non-aggressiveness--are almost certainly polygenic to the extent that they have genetic underpinnings, it follows that similar breeding schemes which incorporate sanctioned programs of AID or embryo transfer, would not be long in coming.

c. Negative Eugenics

Negative eugenics applies to policies and program with the intended result of reducing the frequency or occurrence of genetically determined disease. In this sense, we now routinely realize "negative eugenic" objectives through amniocentesis and selective abortion of genetically handicapped fetuses, even though some of these steps will in fact have dysgenic impacts (e.g., as the previously discussed result of reproductive compensation following abortion of a homozygous recessive individual: see pages 30-31). In this latter instance the effect of couples wishing to "make up for" the missing affected child lost as a result of selective abortion is to increase the relative frequency of the deleterious allele, since 2 out of 3 of their children will be carriers of the deleterious gene, assuming that they continue to abort homozygous affected fetuses.

Negative eugenics also implies the selective elimination of gametes (eg, egg or sperm) which carry deleterious genes, as well as discouraging their carriers from procreation. There are few technical obstacles in place which would prevent the application of gene sequencing studies on humans that would use such principles to achieveve eugenic effects. Specifically, the use of gene sequencing of human sperm designated for AID (aritificial insemination by donor) to exclude samples of dubious genetic quality is actively being considered. An analogous screening process has already been recommended by the Ethics Committee of the American Fertility Society (74).

The availability of the full genetic sequence does not immediately imply that its use for such a screen is technically possible: indeed, it will undoubtedly require new innovations to expand the present gene probe analysis (or related techniques)to the large number of major and minor genetic loci needed to make sperm donor testing feasible. Nonetheless, it is likely to be the first major use of the genomic data with potential eugenic impacts.

Among the reconmmendations for a "minimal genetic screen for gamate donors," is the recommendation that the donor not be found to carry an autosomal recessive gene for any disease known to be prevalent in the donor's ethnic background"(75). Presumably the limitation to "ethnic background" is based on the limited availability and expense of genetic tests. Once a full "gene screen" becomes technically and economically feasible--and such feasibility will be immensely enhanced through the genome project--this limitation will unlikely apply.

d. Nefarious Uses

The most questionable uses of the new genomic information would appear to reside outside the sphere of reproduction, specifically where that data applies to living persons whose impairments limit their ability to be gainfully engaged in society. The most extreme version of such a program would be in the instance where presumptively inherited conditions were cited as the justification for "eugenic sterilization".

While such Nazi-like policies are almost universally seen as morally repugnant, genetic screening to identify individuals with marginal capabilities or handicaps and subsequent programs to isolate tham from society have not been so regarded. The counter argument is that genetic tests will allow us to more accurately define irreversible mental handicap and treat persons accordingly. Thus, we routintely housed and isolated the mentally handicapped in places like Willowbrook, Long Island without benefit of tests which could discriminate persons who were palsied but otherwise intelligent from those who were genuinely intellectually and developmentally limited (83). (Of course, many would question the warehousing of any mentally handicapped child in the first place.

The area of greatest concern focusses on those areas of human ability that provide the basis for culling or selecting students for placement in school or other special services, since in time such policies reinforce selective mating that can have seemingly "eugenic" impacts. Were a "genetic screen" to embrace dyslexia, fragile X syndrome, and manic depressive psychosis, it would have a dramatic impact on educational policy nationwide.

e. "Normal" eugenics

The third area of use of genomic information would be in assuring not merely that the person lacked severe incapacitating genetic conditions, but that he have at least a modicum of "normal" genes. Thus, Bently Glass has posited that individuals have "as the paramount right, that of every child to be born with a normal, adequate hereiditary endowment (76). This prescription presupposes that we will know and define a normal reaction range for a given sub-group of our species. But the idea, once seemingly farfetched, is much closer with the advent of a full genome sequence.

As such it is much more difficult to rebut than by invoking the assertion "we cannot define normalcy" (because of the high frequency of genetic polymorphisms etc), and hence could not apply the technology appropriately. What is needed is a full scale reassessment of the Glass position. If we could define genetic normalcy, would it be something to which each person was entitled?

This notion could in theory be based on establishing the norm of most common genotypes for a given ethnic group in our society and than asking that each child be allowed (via pre-natal diagnosis) to approximate that genotype. Parents would conceivably be offered the information on which to judge the appropriateness of a selective abortion. This precept would violate the belief that persons are entitled to existence irrespective of their genotype, or are entitled to a randomly selected genetic makeup based on the composition of chance arrangements of their parental gametes. Leon Kass as well as the proponents of the Jewish, Protestant and Catholic denominations would probably find the notion of a "normal" human genotype and its selection morally repugnant (cf 77).

But certain, less egregious uses of genetic criteria to define normalcy oreligibility for certain societal goods; or to assign individuals into "appropriate" grades or categories (beyond those stratifications for educational policy previously discussed), are highly likely. For instance, a partially tongue-in cheek rendition of what sports testing will look like in the near future postulated on the basis of our gene mapping and DNA probe technologies:

"...it soon will be feasible to locate, measure and categorize those stretches of the human genome that determine a sprinter's capacity to ferry oxygen to the tissues on the hemoglobin of red blood cells. There assuredly will be differences in this fragment of physiology between one individual and another, and between populations from different parts of the world..."(78).

Others like Nobel Laureate, Sir Macfarlane Burnet concur that genes determine appreciable portions of the phenotype. Burnet has stated that "nearly all the observable differences among people that are not simply related to age, sex, or disease are predominantly based on genetic differences (79). While Burnet appears willing to allow the natural lottery to generate only a few creative individuals each generations, others are not so laissez faire. (Not very long ago, an entrepeneur established a sperm bank exclusively for Nobel laureates on the premise that such "blind" luck could be circumvented by using sperm donors with "proven" attirbutes.

f. Voluntarism

Should the genetic information be used to identify and exclude or otherwise limit the rights and freedoms of already extant individuals, major ethical issues arise. As geneticist Arno Motulsky has observed, "As long as such knowledge only concerns genes affecting variables of physical health and as long as testing remains voluntary, society might be able to cope. But when we learn more about the genetics of personality and mental traits, new problems could arise" (81).

g. Privacy

Most of the protections identified by consensus committees such as the NAS, President's Commission and the Hastings Center Genetics Research Group focus on protective procedures for safeguarding genetic knowledge (e.g., informed consent, protection of privacy and confidentiality). But at least one such committee, the US NAS foresaw the need to override individual interests in safeguarding genetic data when others might be placed at risk by not knowing. According the the NAS Committee, it was foreseeable that a need would arise "to consider statutes permitting the conveying of genetic information to persons at risk , without their request and without the permission of the person from whom the information is derived."(82).

While such a provision was initially intended to be limted to information "that would alert a persn to a life-threatening or massively disabling condition", it can be seen as inviting a wide interpretation as to what constitutes "life-threatening" or "disabling". A workplace screener might justifiably invoke such a declaration in asking his workers to submit to compulsory testing for catastrophic (albeit extremely rare) conditions that have a genetic basis such as succinyl choline esterase and malignant hyperthermia (e.g., among anesthesiologists) or the more common G6PD and hemolysis (e.g., among naptha workers). A balance point between individual and societal rights could be drawn at the point where risk of harm to others outweighed the right to privacy of the individual worker.

Here the clash between the deontological and utilitarian traditions is most evident: To some ethicists, workplace screening for hypersusceptible workers would be acceptable if it included safeguards, while for others, no testing would be permissable because of the prospect of further abuses once the principle were allowed (cf 83, 84).

7. Issues of Control of Knowledge

The penultimate value of the genetic knowledge which will be gathered through the genome project will not likely be realized until some decades after its assembly. There is little doubt that such a data base will have great utility initially in helping elucidate the basis and physical location of the genes that cause or predispose to illness and disease. As such, the sequences themselves (and earlier the map and linkage groups) will have substantial commercial value.

This commercial value has already been recognized by major companies such as Collaborative Research of Bedford, Massachusetts which has identified over 60 markers on Chromosome 7 alone, and has applied for patents on a number of specific gene probes. Along with Walter Gilbert, Collaborative and other gene sequencing companies have argued for their intrinsic right to patent and control the data set they develop.

The net result of the selective forces of the marketplace has been to generate a database network, in which individual companies hold various portions of the sequence story. The largest is the federally supported GenBank sponsored by NIH. Smaller agencies such as the American Type Culture Collection have their own set of probes which they will sell to researchers both in the private and public sector at reasonable rates (i.e. around $100/probe)(85).

The ethical issues of privatization of this knowledge turn on the importance of sequences lost to the academic community by corporations who have kept their own sequences secret. On one level the problem is largely academic, since the sequences needed for a complete picture of human DNA could be assembled by the public sector, with duplication of the sequences held secret by private parties. On another, the potential loss of critical data, the dukplicaion of effort and the sequestration of knowledge raise serious questions of public versus proprietary holding of fundamental knowledge.

a. Options: The options of the government appear at least threefold:

1. Permit the status quo, allowing sequestration of some knowledge with the assurance that patenting will eventually permit the free flow of most crucial data.

2. Exert control over that data which is generated from governmentally funded sources, while allowing the private sector to continue sequencing work.

3. Exert control over both areas of inquiry, insisting that sequence information is a public resource.

What constitutes appropriate "control" is itself an ethically fraught issue. Patenting is an instance of "control". So is publication in a copyrighted journal. The government may reserve the right to limit the flow of information to the public in the public interest. Or, it may insist on the full disclosure and recording of sequences (in say GenBank) as a condition for federal support. Certain sequences, for instance those indicating the existence of a genetic predisposition to retardation, might be protected in one way, while others such as those linked to specific hereditary diseases another. It would be entirely acceptable ethically to have the federal government decide that its level of control would be over protecting individually identifiable sequences only, with all other sequences to be filed in GenBank and thereby be made publicly accessible.

Under the first model, it must be noted that the law allows the patenting of specific sequences, particularly if they are incorporated into a probe, on the ground that as purified substances with a novel function they constitute "new uses" (i.e. not previously found in nature) of a biological substance. From the perspective of societal value, a patenting process of some kind ensures the public disclosure (or derivability) of a given sequence, while maintaining an incentive to promulgate the knowledge through rewards (e.g., exclusive licensing, etc.) to the inventor.

Any newfound utility "in the hand of man" potentially blurs the distinction between sequences acquired for specific hypothesis testing, as fundamental knowledge, or as applied diagnostic or therapeutic aids. In practice, the first policy might well result in the sequestration of much more knowledge than that encased in probe sequences alone.

From the ethical perspective of individual rights supporters like Robert Nozick, the privatization of this knowledge can be condoned as long as the sequestration of data makes the public no worse off than they would without it. In this sense, a gene sequence deserves protection as an element of intellectual property rights. Others have countered that no one has the right to patent truly basic knowledge: ideas per se have never been granted patentability (86).

b. Public interest

The principle of public interest postulates that the degree of governmental involvement and control should be porpotional to the capacity of a technology to impact on public welfare and individual well-being. It is in part for these reasons that Alan Goldman of the university of Miami has argued that publicly sponsored and controlled research should remand to the public (86). From Goldman's position, it would appear that the public deserves access to and the rewards from any and all data from research projects like those entailed in the genome project, particularly since this area and its findings related directly to public health.

Others, specifically Ira Carmen of the University of Illinois at Urbana/Champaign, have effectively argued that the DNA sequence lies in the marketplace of ideas and hence is protected under the First Amendment (87). According to Carmen, because Congress cannot limit the dissemination of views of the biological world, it cannot delimit the marketplace of genome sequencing by granting the right of access to a portion of that data. What this analysis misses is the prior question of why if at all the human genome comprise an "idea marketplace".

c. Arguments

It would appear to be a fundamental truth that the human genome is a collective property, held in common among all persons of human heritage. The fact that some people will have ("own") a unique sub-fraction of this total (e.g., those who have the mutant gene for alpha 1 antitrypsin) does not vitiate this claim. While the argument that the human genome is already "in the public domain" may not be legally sound, it has a ring of moral truth to it. We do have special rules for "human" property as opposed to all other kinds: we do not allow children to be bought or sold; we limit how and under what circumstances contracts may be written for embryos fertilized by other than natural means; and we have special rules for human cell lines and cultures (87).

It would be ethically congruent with the special standing given to protections of like human materials, to afford special status to human genesequences, especially those which are held in common among all or most persons. Patenting of a genetically unique cell line, by this reasoning, would be qualititatively different than patenting a commonly held DNA sequence as a probe, or copyrighting its promulgation.

A countervailing argument is that since gene sequences are not commonly "knowable" and require the use of expensive and often uniquely patented machinery to elucidate, that their discovery belongs to the person(s) who uncovered them. According to this reasoning, it does not matter how the sequences might be used or whether or not they are unique: rather it is the fact of the labor and invention invested in their discovery which makes them valid intellectual property.

8. Summary and Recommendations

The likelihood that the human genome project will be of inestimable value for both basic and applied research dictates that special conditions and terms be set for assuring the most efficient and equitable acquisition of the necessary data. Because of the intrinsic capacity of genomic information to be used to discriminate (both invidiously and acceptably) among members of the population and to affect their entitlements, it is also highly desirable that protections against significant abuse be in place before the project gets underway.

To maximize the efficient and equitable acquisition of data and afford reasonable protections to their use, Congress should consider implementing one or more of the following policies:

1. Assign the genome project status as a national resource, protecting both those who acquire the data (e.g., by offering idemnification from harms), and the resulting knowledge base (e.g., by encoding the sequences in a national archive analogous to the GenBank system already in place) with special safeguards).
2. Protect the privacy of individually identifiable data.
3. Take steps to place the genome registry in the public domain.
4. Offer safeguards against eugenic uses of the data which go beyond

identifying deleterious genes associated with serious human disorders or disease.

5. Permit responsible use of DNA sequence data in probes designed to identify genetic disorders.

Because of the likelihood that the project will generate data of unforeseeable consequences, a special Human Genome Project Oversight Committee should be put in place during the project's duration to monitor its progress and to anticipate both scientific and ethical implications of its findings.

REFERENCES

1. Anonymous, "Mapping the human genome," Lancet 1: 1121-22, 1987.

2. Anonymous, "DuPont's DNA sequencer uses new chemistry," Bio/Technology 5: 1111-12, 1987.

3. Charles R. Scriver, et al, "Genetics and medicine: An evolving relationship," In: Health Care, Philip Abeslon, editor, AAAS (Washington, DC 1978).

4. P.H. StGeorge-Hyslop, R.E. Tanzi, R.J. Polinsky et al, "Absence of duplication of chromosome 21 genes in familial and aporadic Alzheimer's disease," Science 238: 664-65, 1987.

5. David Tyrrell, The Epstein-Barr Virus, M.A. Epstein and B.G. Achong, eds (Heinemann, London, 1986), Book Review, Lancet 2: 771, 1987.

6. D.D. Kostyu and E.G. Reisner, "Human leukocyte and platelet antigens and antibodies," In: Hematology, 3rd Edition, W.J. Williams et al, editors, Mc Graw Hill (New York, 1983).

7. Richard F. Selden, Marek J. Skoskiewicz, Paul S. Russell, and Howard M. Goodman, "Regulation of insulin-gene expression: Implications for gene therapy," New England Journal of Medicine 317: 1067-1076, 1987.

8. Genetic prediction and family structure in Huntington's chorea," British Medical Journal 290: 1929031, 1987.

9. M.E. Eyster, J.J. Goedert, M-C Poon, and O.T. Preble, "Acid-labile alpha interferon: A possible preclinical marker for acquired immunodeficiency syndrome in hemophilia," New England Journal of Medicine 309: 583-6, 1983.

10. Jeremy Rifkin, "One small step beyond mankind," Progressive 41: 21, March, 1977.

11. President's Commission for the Study of Ethical Problems in Medicine and

Biomedical and Behavioral Research, *Screening and Counseling for Genetic Conditions: A Report on the Ethical, Social, and Legal Implications of Genetic Screening, Counseling and Education Programs*, U.S. Gov't Printing Office (Washington, DC 1983), see especially pp 41-86.

12. Horace Freeland Judson, "Fearful of science," *Harpers* 25: 36-41, (March 1975), 70-76, (June, 1975).

13. Robert Jay Lifton, *The Nazi Doctors: Medical Killing and the Psychology of Genocide*, Basic Books, (New York, 1986).

14. George L. Mosse, *Masses and Man: Nationalist and Fascist Perceptions of Reality*, Fertig (New York, 1980).

15. Daniel J. Kevles, *In the Name of Eugenics: Genetics and the Uses of Human Heredity,* University of California Press (Berkeley, 1986).

16. Marc Lappe, et al, "Ethical and social issues in screening for genetic disease, *New England Journal of Medicine* 286: 1129-1134, 1972.

17. Elof A. Carlson, *Genes, Radiation, and Society: The Life and Work of H.J. Muller*, Cornell University Press (Ithaca, NY 1981), see especially pp 186-90.

18. Winnifred Gallagher, "To the manner born," Rolling Stone, Nov 19, 1987, p. 56 ff.

19. R.B. Campbell, "The effects of genetic screening and assortative mating on lethal recessive-allele frequencies and homozygote incidence," *American Journal of Human Genetics* 41: 671-77, 1987.

20. Reed E. Pyerita, Jane E. Tumpson and Barbara A. Bernhardt, "The economics of clinical genetics services. I. Preview," *American Journal of Human Genetics* 41: 549-558, 1987.

21. Ira M. Rutkow and Jeffrey M. Lipton, "Mandatory screening for sickle cell anemia," *New England Journal of Medicine* 289: 865, 1973 (letter); see also

"Grounded for Bearing Sickle Cell Trait, Washington Post, Deceember 26, 1980, p A9.

22. Haig H. Kazazian, "Prenatal diagnosis for sex choice: A medical view," Hastings Center Report 10: 17-19, 1980.

23. D.W. Cox and T. Mansfield, "Prenatal diagnosis for alpha-1-antitrypsin deficiency," Lancet 1: 230-232, 1985.

24. D.L. Hull, "Discussion: Reduction in genetics," Philosophy of Science 46: 316-20, 1979.

25. Patricia Kitcher, "Discussion: Genetics, reducton, and functional psychology," Philsophy of Science 49: 633-36, 1982.

26. Arthur L. Capalan, "Why the problem of reductionism in biological science will not go away," Growth 51: 22-34, 1987.

27. Harold M. Schmeck, Jr, "Potent tool fashioned to probe inherited ills," New York Times August 11, 1987, pp 19, 22.

28. Stanley S. Robin, Gerald E. Markle et al,"...Let no one split asunder: Controversy in HUman Genetic Engineering," Politics and the Life Sciences 6: 3-26, 1987.

29. See G. M. Wickens, The Nasirean Ethics by Nasir al-Din Tusi, UNESCO Collection of Representative Works: Persian series, Allen & Unwin, (London, 1964); see also, S. Salmon Nadvi, "Preliminary observations form the Islamic point of view," In Genetics and Society, pp. 131-2.

30. Paul Ramsey, Fabricated Man: The Ethics of Genetic Control, Yale University Press (New Haven, 1970), see especially pp 29-30.

31. Michael D. Bayles, Reproductive Ethics, Prentice Hall (New Jersey, 1984).

32. S. Karlin and P. O'Donald, "Some population genetic models combining sexual selection and assortative mating," Heredity 41: 165-74, 1978.

33. Richard McCormick, "Dissent on the use of third parties," Appendix A In: Ethical Considerations of the New Reproductive Technologies, *Fertility and Sterility*, Supplement 1 46: 82S, 1986.

34."Text of the Vatican Doctrinal Statement on Human Reproduction," *New York Times*, March 11, 1987, pp A14-A17.

35. Immanuel Jakobovits, "Some letters on Jewish medical ethics," *Journal of Medicine and Philosophy* 8: 217-224, 1983.

36. National Council of Churches of Christ in the U.S.A. Panel on Bioethical Concerns, *Genetic Engineering: Social and Ethical Consequences,* Pilgrim Press (New York, 1984); see also, *Genetic Science for Human Benefit*, National Council of Churches (New York, 1986).

37. Jeremy Rifkin, *The Tehological Letter Concerning the Moral Arguments Against Genetic Engineering of the Human Germline Cells*, Unpublished document available from the Foundation on Economic Trends, 1346 Connecticut AVE NW, Suite 1010, Washington, DC 20036, June 1983, 13 pages.

38. Jacques Ellul, *The Technological Society*, Vintage Press (New York, 1964), p 11 et seq.

39. Leon R. Kass, "The New Biology: What Price Relieving Man's Estate?", In: Toiward a More Natural Science: Biology and Human Affairs, The Free Press (New York, 1986), p. 33.

40. Magdalena Ruiz, "Rules on embryos proposed," *The Scientist*, September 7, 1987, p. 5.

41. Joseph Fletecher, "Ethical aspects of genetic controls," *New England Journal of Medicine* 285: 776-83, 1971.

42. Leon R. Kass, "Perfect Babies: Prenatal Diagnosis and the Equal Right to Life," In: *Toward a More Natural Science*, p. 80-98.

43. Jeremy Rifkin, cited in "Briths", Chicago Tribune, October 19, 1987, p. 10.

44. Daniel Callahan, "What obligations do we have to future generations?", American Ecclesiastical Review 164: 265-280, 1971.

45. Martin P. Golding, "Obligations to future generations," Monist 56: 85-99, 1972.

46. President's Commission, Securing Access to Health Care: The Ethical Implications of Differences in the Availability of Health Services, Volume 1, Report, U.S. Gov't Printing Office (Washington, DC 1983), pp 11-48.

47. T.D. Bird, "Medical genetics and clinical neurology: The common ground," Progress in Medical Genetics 6: 1-18, 1985.

48. William K. Frankena, Ethics, 2nd Edition, Prentice-Hall, (Englewood Cliffs, NJ, 1963).

49. Jessica G. Davis, "A Counselor's Viewpoint," Genetic Counseling: Facts Values and Norms, Birth Defects Original Article Series 15: 113-122, 1979.

50. Mark A. Rothstein, "Discriminatory aspects of medical screening," Journal of Occupational Medicine 28: 924-29, 1986.

51. D.C. Wertrz and J.R. Sorenson, "Client reactions to genetic counseling: self-reports of influence," Clinical Genetics 30: 494-502, 1986.

52. James R. Sorenson and Arthur J. Culbert, "Professional orientations to contemporary genetic counseling," Birth Defects Orginal Article Series 15: 85-102, 1979.

53..L. Smith and L. Hood, "Mapping and sequencing the human genome: how to proceed," Bio/Technology 5: 933-939, 1987.

54. Robert A. Weinberg, "The case against gene sequencing," The Scientist, November 16, 1987, p. 11.

55. Lawrence Hall of Science, "Questions: Project in Understanding Science in

Secondary Education, Personal Communication to M. Lappe, October 20, 1987.

56. John R. Platt, "Strong inference," *Science* 146: 347-352, 1964.

57. Daniel Callahan, "Ethical responsibility in science in the face of uncertain consequences," *Annals of the NY Academy of Sciences* 265: 1-12, 1976.

58. Constance Holden, "The genetics of personality," *Science* 237: 598-601, 1987.

59. Herman Kahn, interviewed by G.R. Urban, "Hereman Kahn thinks about the thinkable:' Most of the traditional causes of war have disappeared'," *New York Times Magazine* June 20, 1971, p 24.

60. Liebe Calvalieri, "New strains of life--or death," *New York Times Magazine*, August 22, 1976, p. 8 et seq.61. Carl Cohen, "When may research be stopped?", *New England Journal of Medicine* 296: 1203-10, 1977.

62. President's Commission, *Screening and Counseling for Genetic Conditions: The Ethical, Social, and Legal Implications fo Genetic Screening, Counseling and Education Programs*, U.S. Govt Printing Office (Washington, DC 1983).

63. Committee for the Study of Inborn Errors of Metabolism, *Genetic Screening: Programs, Principles and Research*, NAS (Washington, DC 1975).

64. Francis Galton, Inquiries into Human Faculty, Macmillan (London, 1883), pp 24-5.

65. Rene Dubos, "Genetic constitution and environmental conditioning," In: Encyclopedia of Bioethics, Vol 2, Warren Reich editor, Macmillan, The Free Press (New York, 1978), pp 548-555.

66. Julian Huxley, "The vital importance of eugenics," *Harpers Monthly* 163: 324-331, August, 1941.

67. Selwyn Franklin, "A preliminary study into the attitude of Judaism," In: G.C. Oosthuizen, H.A. Shapiro and S.A. Strauss, editors Genetics and Society,

Oxford University Press (New York, 1980)

68. Immanuel Jakobovits, *Jewish Medical Ethics: A Comparative and Historical Study of the Jewish Religious Attitutde to Medicine and its Practice*, New Edition, Bloch (New York, 1975).

69. ________________, Eugenics and Religious Law: Jewish Religious Laws, In: Encyclopedia of Bioethics, Vol 1, pp 468-470.

70. **William** W. Bassett, "Christian Religious Laws", In: *Encyclopedia of Bioethics*, Vol 1 pp 471-473.

71. The President's Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, *Splicing Life: A Report on the Social and Ethical Issues of Genetic Engineering in Human Beings*, U.S. Govt Printing Office, (Washington, DC 1982).

72. C.K. Chan, "Eugenics on the rise: A report from Singapore," International Journal of Health Services 15: 707-713, 1985.

73. J.S. Beckman, "Molecular markers in the genetic improvement of farm animals," *Bio//Technology* 5: 573-78, 1987.

74. Ethics Committee of the American Fertility Society, "Ethical Considerations of the New Reproductive Technologies", *Fertility and Sterility Supplement 1* 46: 1S-84S, 1986.

75. *Ibid.*, Appendix B, pp 83S-84S.

76. Bentley Glass, "Ethical problems raised by genetics," In: Genetics and the Quality of Life, Charles Birch and Paul Abrecht editors, Pergamon Press (New Yorik, 1974), pp 51-57.

77. Leon R. Kass, "Looking Good: Nature and Nobility," In: *Toward a More Natural Science*, op cit., pp 318-345.

78. Bernard Dixon, "What's the sporting use of science," *The Scientist*,

October 19, 1987, p. 13.

79. F.M. Burnet, "Man and his genes," *Perspectives in Biology and Medicine*, Summer, 1978, pp 505-514.

80. J. Robertson, "Procreative liberty and the control of conception, pregnancy and childbirth," *Virginia Law Review* 69: 405-432, 1983.

81. Arno Motulsky, "Brave New World? Current approaches to prevention, treatment and research of genetic diseases raise ethical issues," *Science* 185: 653-6, 1974.

82. Committee on Genetics, "Legal principles for genetic screening," In: *Genetic Screening: Programs Principles and Research*, National Academy of Sciences, (Washington, DC 1978) pp 183-193.

83. Gordon Atherly, Neil Johnston and Maritza Tennassee, "Biomedical surveillance: rights conflict with rights," *Journal of Occupational Medicine* 28: 958-961, 1986.

84. Marc A. Lappe, "Ethical concerns in occupational screening programs," *Journal of Occupational Medicine* 28: 930-934, 1986.

85. Leslie Roberts, "Who owns the human genome?', *Science* 237: 358-361, 1987.

86. Ira H. Carmen, "Ownership of the human genome," Science 237: 1555, 1987.

87. Great Britain Department of Health and Social Security: *Report of the Committee of Inquiry into Humnan Fertilisation and Embryology*, M. Warnoick Editor, Her Majesty's Stationery Office, (London, 1984).